

AI-Driven Image Generation: Algorithms, Architectures, Quality Assessment, and Applications—A Structured Narrative Review

Mikołaj Leszczuk ^{1,*} , Yi Zhang ² , Mylène C. Q. Farias ³ , Damon M. Chandler ⁴  and Ruth Kalola ²

¹ AGH University of Krakow, 30-059 Kraków, Poland

² Xi'an Jiaotong University, Xi'an 710049, China; yi.zhang.osu@xjtu.edu.cn (Y.Z.); ruthkalola@stu.xjtu.edu.cn (R.K.)

³ Texas State University, San Marcos, TX 78666, USA; mylene@ieee.org

⁴ Ritsumeikan University, Osaka 567-8570, Japan

* Correspondence: mikolaj.leszczuk@agh.edu.pl; Tel.: +48-607-720-398

Abstract

Background: This paper presents a structured narrative review of recent advances in AI-driven image generation across four complementary perspectives: generative models and architectures, quality assessment and performance metrics, application domains, and multi-modal or cross-lingual extensions. The review aimed to identify dominant methodological trends, representative evaluation practices, and open research challenges in contemporary image generation research. **Methods:** A structured literature search was conducted in the Scopus database on 29 January 2026 using a predefined query focused on modern generative-image paradigms and excluding clearly out-of-scope domains. Eligible records addressed contemporary AI-driven image generation or closely related multi-modal generation settings within the temporal and topical scope of the search. Retrieved records were first assigned to four thematic branches and then screened with branch-specific relevance criteria for narrative synthesis. **Results:** The search returned 1524 records, and the final narrative synthesis included 117 publications: 34 on generative models, 23 on quality assessment, 29 on applications, and 31 on multi-modal and cross-lingual aspects. Across the reviewed literature, progress was shaped not only by visual fidelity, but also by controllability, semantic grounding, human-centred evaluation, multi-modal integration, and practical deployment constraints. **Limitations:** The review was limited to a single primary bibliographic source and to a qualitative narrative synthesis without meta-analysis. **Conclusions:** The review provides a structured reference point for researchers and practitioners working on AI-based image generation and its evaluation, while also highlighting benchmark, comparability, and multi-modal-transfer challenges that remain unresolved.

Keywords: AI-driven image generation; diffusion models; generative adversarial networks; foundation models; text-to-image synthesis; image quality assessment; multi-modal learning; structured narrative review



Academic Editor: Silvia Liberata Ullo

Received: 29 March 2026

Revised: 25 June 2026

Accepted: 27 June 2026

Published: 1 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Recent advances in artificial intelligence have significantly accelerated the development of generative models capable of producing high-quality synthetic visual content. In particular, the emergence of diffusion models, generative adversarial networks (GANs), and large-scale transformer-based foundation models has enabled substantial progress in text-to-image generation and prompt-driven visual synthesis. These systems allow users to

generate complex images directly from textual descriptions, enabling applications ranging from digital art and design to scientific visualization and content generation.

Early approaches to image generation were primarily based on generative adversarial networks, which demonstrated the feasibility of synthesizing visually realistic images through adversarial training. However, recent diffusion-based architectures and large-scale multi-modal foundation models have significantly improved both the quality and controllability of generated images. Modern text-to-image systems integrate natural language understanding with visual generation pipelines, enabling increasingly accurate alignment between user prompts and generated content.

Despite these advances, several challenges remain in the development and evaluation of AI-generated imagery. These challenges include ensuring semantic alignment between textual prompts and visual outputs, maintaining perceptual realism, evaluating generated images using reliable objective metrics, and integrating multi-modal reasoning within generative architectures. Furthermore, the rapid growth of the field has resulted in a fragmented research landscape, with contributions spanning generative model architectures, multi-modal fusion strategies, evaluation methodologies, and application-specific deployments.

To better understand the evolution of this rapidly developing domain, this article presents a structured review of recent research on AI-driven image generation. The review focuses on four major research directions: (1) generative model architectures enabling modern text-to-image synthesis, (2) evaluation frameworks and quality assessment metrics for AI-generated images, (3) emerging applications of generative image models across different domains, and (4) multi-modal representation learning and cross-modal alignment techniques.

Research Questions and Objectives

This structured narrative review is structured around the following four primary research questions:

- (1) **Architectures and Design Paradigms:** Which architectural paradigms and design strategies dominate current AI-driven image generation research, and what are the key innovations in model families such as GANs, diffusion models, and foundation models?
- (2) **Quality Assessment and Evaluation:** How are the quality, realism, semantic faithfulness, and perceptual alignment of AI-generated images evaluated across the literature, and which objective and subjective metrics are most commonly employed?
- (3) **Application Domains:** Which application domains have emerged as substantively studied in recent literature, and what are the primary use cases and deployment contexts for AI-driven image generation?
- (4) **Multi-modal and Cross-Lingual Integration:** How are multi-modal and cross-lingual mechanisms being incorporated into contemporary image-generation workflows, and what role do they play in enhancing semantic alignment and generation quality?

More specifically, the review addresses the following questions: (1) which architectural paradigms and design strategies dominate current AI-driven image generation research; (2) how the quality, realism, and semantic faithfulness of generated images are evaluated; (3) which application domains have emerged as the most substantively studied in the recent literature; and (4) how multi-modal and cross-lingual mechanisms are being incorporated into contemporary image-generation workflows. These questions define the analytical scope of the review and motivate the four-part thematic structure adopted in the remainder of the manuscript.

Figure 1 illustrates the conceptual taxonomy of the main research directions analyzed in this structured narrative review, including generative architectures, evaluation frameworks, application-oriented deployments, and multi-modal modeling strategies for AI-driven image generation.

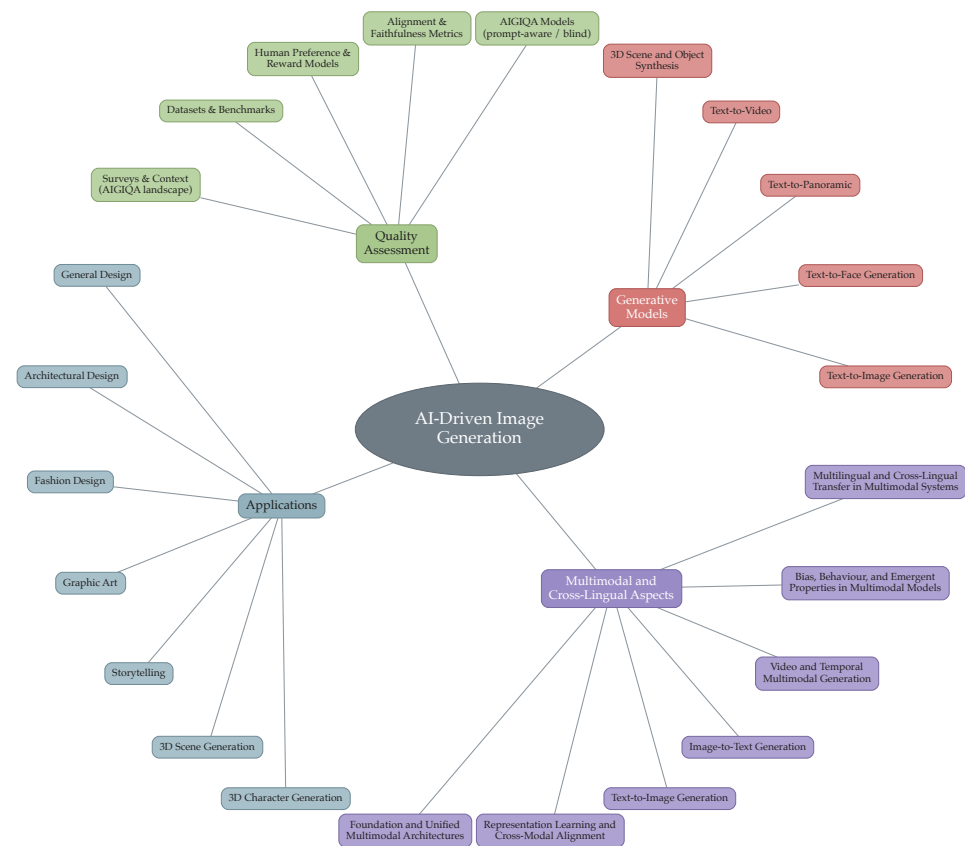


Figure 1. Conceptual taxonomy of the main research directions covered in this structured narrative review.

The reviewed literature indicates a rapid expansion of research on image generation using modern deep learning paradigms, including GANs, diffusion models, transformers, and foundation models. Across the analyzed corpus, clear trends can be observed in methodological development, thematic diversification, and the growing breadth of application areas.

Figure 2 illustrates the temporal distribution of the collected publications and highlights the rapid increase in research activity following the emergence of large-scale diffusion models and multi-modal foundation models. The observed trends indicate a clear shift from early GAN-based approaches toward diffusion-driven architectures and multi-modal generative frameworks that integrate language understanding with visual synthesis.

The remainder of the paper is structured as follows. Section 2 describes the systematic literature search and selection methodology used to construct the analyzed publication corpus. Subsequent sections analyze the literature from four complementary perspectives: generative model architectures, multi-modal modeling strategies, evaluation and quality assessment metrics, and application-driven deployments of AI-generated imagery.

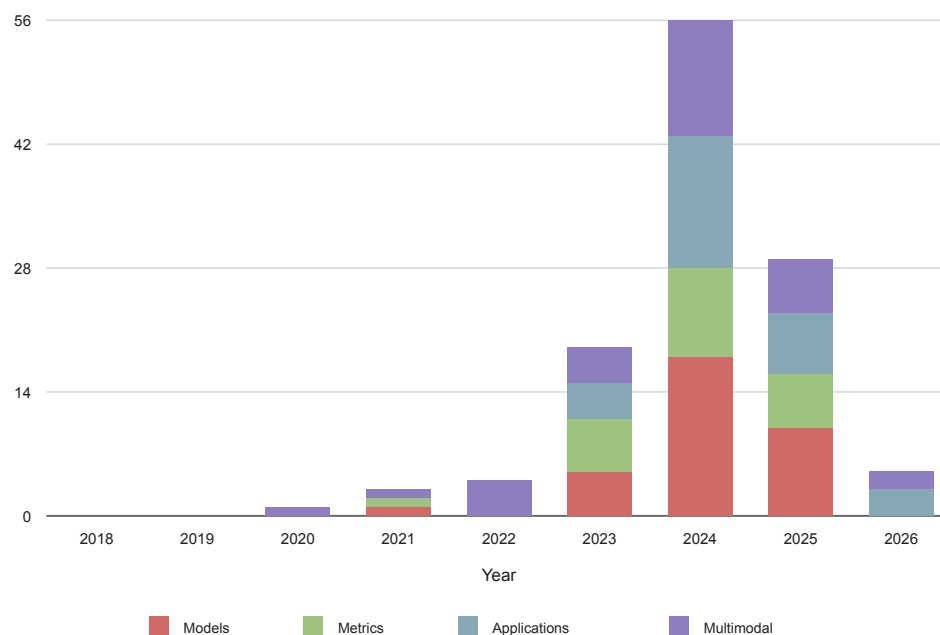


Figure 2. Timeline and distribution of publications analyzed in this structured narrative review of AI-driven image generation research.

2. Methodology

This study followed a structured synthesis workflow designed to ensure transparency, reproducibility, and balanced coverage across the major research directions in AI-driven image generation. The review was reported with reference to the PRISMA 2020 guideline for systematic reviews, adapted to a qualitative narrative synthesis rather than a meta-analysis. The literature collection process was conducted using the Scopus database on 29 January 2026. Scopus was used as the primary information source for record identification in order to provide a broad and internally consistent index of peer-reviewed literature across computer science, engineering, and applied AI venues. Extending the search to additional bibliographic databases such as Web of Science or IEEE Xplore was considered during the design phase; however, for the specific publication types and venue terminology characteristic of the AI image generation domain, Scopus provides extensive indexed coverage of the relevant conference proceedings and journals, and the additional retrieval gain from a parallel search in secondary sources was judged unlikely to change the composition of the final synthesis corpus materially. The single-source design is nevertheless acknowledged as a coverage constraint and is reported explicitly as a transparency limitation in the Discussion section. To avoid omitting a small number of field-defining systems that did not yet have a clearly identifiable peer-reviewed counterpart at the time of revision, the final narrative synthesis also retained a limited number of directly relevant public technical reports or arXiv manuscripts, which are explicitly identifiable as such in the reference list. For the purposes of this review, a publication is treated as field-defining if it introduced a generative architecture or evaluation framework that is widely cited as a foundational baseline in the subsequent peer-reviewed literature, irrespective of whether the source document itself underwent formal peer review. The search query was intentionally formulated to capture publications focused on modern image generation paradigms, including GAN-based, diffusion-based, transformer-based, and foundation-model approaches, while excluding domains not directly relevant to prompt-driven visual synthesis, such as restoration, denoising, image-to-image translation, medical imaging, and remote sensing.

The exact Scopus query used in the study was as follows:

```
( 'image generation' ) AND
( GAN OR diffusion ) AND
( transformer OR 'foundation model' ) AND
( 'text-to-image' OR prompt-based OR prompt-driven ) AND NOT
( augmentation OR enhancement OR restoration OR denoising
  OR 'image-to-image' OR 'image translation'
  OR medical OR biomedical OR radiology OR 'remote sensing' )
AND PUBYEAR > 2017 AND PUBYEAR < 2027
```

This deliberately conjunctive (AND-only) query design was selected to privilege topical specificity over broad recall and to keep the initial corpus focused on modern prompt-driven image-generation paradigms. The trade-off is that some otherwise relevant studies using only adjacent terminology (without explicitly matching all core concept blocks) may not be retrieved at the database-query stage. This limitation is considered in the interpretation of coverage and is reported explicitly to preserve transparency.

2.1. Eligibility Criteria

Inclusion Criteria:

- Contemporary AI-driven image generation (e.g., GANs, diffusion models, transformers, foundation models) or closely related multi-modal generation settings.
- Publication year between 2018 and 2026 (consistent with Scopus query temporal scope).
- Peer-reviewed or field-defining public technical reports and arXiv manuscripts in AI image generation.
- Methodological, evaluative, or application-oriented contribution relevant to one of the four review branches: (1) generative models and architectures, (2) quality assessment and metrics, (3) applications, or (4) multi-modal and cross-lingual aspects.

Exclusion Criteria:

- Image restoration, denoising, super-resolution, inpainting, image-to-image translation, style transfer.
- Medical imaging, biomedical imaging, radiology, remote sensing, satellite imagery.
- Deepfake detection, forensic analysis, watermarking, copyright and legal analysis.
- General image processing, computer vision, or perception without explicit generative focus.
- Unrelated optimization, training techniques, or infrastructure papers.

The query returned a total of 1524 records. The review eligibility criteria were defined before branch-level screening and applied consistently across the corpus. Inclusion required that a study: (i) addressed contemporary AI-driven image generation or a closely related multi-modal generation setting; (ii) fell within the temporal and topical scope of the Scopus query; and (iii) contributed methodological, evaluative, or application-oriented evidence relevant to one of the four review branches. Exclusion applied to studies located in explicitly excluded domains or clearly centred on tasks outside the intended review scope, such as restoration, denoising, image-to-image translation, medical imaging, remote sensing, or unrelated image-processing pipelines. Rather than screening the entire corpus manually as a single undifferentiated pool, the retrieved papers were first assigned to thematic categories using a transparent rule-based keyword matching scheme applied to titles, abstracts, and author keywords. Each record was scored against four category-specific keyword lists. For each category, the score was computed as the aggregate frequency of matched keywords (after whitespace normalization and case-insensitive matching) appearing in the title, abstract, and author keyword fields. A record was assigned to the category with the highest aggregate score. In the event of a tie, the lower-numbered category received priority.

The four category keyword sets used for assignment were: (1) Generative Models and Architectures—*architecture, network, model, gan, diffusion, transformer, unet, latent, training, scalability, foundation model, text-to-image model*; (2) Quality Assessment and Performance Metrics—*quality, evaluation, metric, assessment, mos, subjective, objective, benchmark, perceptual, user study, human study, fidelity*; (3) Applications—*application, applied, use case, deployment, industry, creative, art, design, game, education, entertainment, content creation*; (4) multi-modal and Cross-Lingual Aspects—*multi-modal, cross-lingual, cross language, text–image, vision–language, vlm, prompt, caption, language, semantic alignment*. Soft balancing was applied retrospectively to verify that no single branch contained a disproportionate share of clearly misclassified records and to resolve any systematic bias introduced by overlapping terminology. This assignment yielded four top-level thematic groups: Generative Models and Architectures for Image Synthesis ($n = 606$), Quality Assessment and Performance Metrics for AI-Generated Images ($n = 304$), Applications of AI-Driven Image Generation ($n = 308$), and multi-modal and Cross-Lingual Aspects of AI-Generated Scientific Content ($n = 306$).

Eligibility decisions were made in two stages. First, records were assigned to one of the four review branches through the above rule-based thematic classification. Second, each branch was screened independently using branch-specific relevance criteria. This stricter stage relied on title-, abstract-, and keyword-level inspection, followed where necessary by closer semantic review of borderline cases. At this branch-specific level, studies were retained only when their primary contribution aligned with the substantive focus of the target section rather than merely mentioning related terminology. Papers whose connection to a branch was incidental, peripheral, or dominated by a different research objective were excluded during this refinement stage. After the initial thematic assignment, section-level screening decisions were made by the responsible branch lead—a designated co-author with domain expertise in the target section topic. Each branch lead independently applied the branch-specific relevance criteria to the assigned record pool. Borderline cases, defined as records whose primary contribution was ambiguous with respect to the branch focus, were resolved through direct discussion between the branch lead and the corresponding author prior to finalising the section corpus. No automation-assisted conflict resolver was implemented; instead, the workflow combined explicit query constraints, rule-based thematic assignment, section-level semantic refinement, and co-author adjudication for borderline cases. This process resulted in final section corpora of 34 papers for generative models, 23 papers for quality assessment, 29 papers for applications, and 31 papers for multi-modal and cross-lingual aspects, giving a final narrative synthesis corpus of 117 publications.

2.2. Data Extraction Protocol and Data Items

For each retained record, an explicit set of data items was defined to standardize information collection:

1. **Bibliographic Metadata:** Authors, publication year, venue (conference or journal), publication type (peer-reviewed vs. preprint), and DOI (where available).
2. **Model Family or Generative Paradigm:** Primary generative architecture (e.g., GAN, diffusion model, VAE, transformer, foundation model, or hybrid approach) and key methodological innovations.
3. **Target Modality and Output Setting:** Input modality (text, image, sketch, etc.), output modality (single image, video, 3D), and multi-modal relationships (if applicable).
4. **Evaluation Context:** Benchmark datasets, evaluation metrics (objective and subjective), user studies, or comparative analyses reported.
5. **Branch-Specific Descriptors:** Application domain, task type, or thematic cluster (architecture family, metric type, application sector, or multi-modal strategy).

6. **Synthesis Role:** Classification of the study's contribution to the review (e.g., methodological advance, benchmark proposal, application case study, or comparative evaluation).

Data extraction was performed within the corresponding section workflow after branch-level selection. Where specific study details were missing, weakly specified, or not directly comparable across papers, the synthesis retained only information explicit in the source article, avoided reconstructing unreported values, and treated the unresolved detail as a qualitative limitation in section-level interpretation. No study investigators were contacted for additional information, and no quantitative outcome extraction framework for meta-analysis was employed.

2.3. Synthesis Methods and Approach

A structured **narrative synthesis** strategy was employed to synthesize the evidence across the four review branches. The approach is characterized as follows:

- **Rationale:** Narrative synthesis was chosen because the included studies were highly heterogeneous in design, task definition, evaluation protocol, and reported outputs. This heterogeneity makes formal statistical pooling and meta-analysis inappropriate for the present review. Narrative synthesis allows for a nuanced, theme-based integration of qualitative and quantitative evidence.
- **Subtheme Organization:** Within each of the four review branches, retained papers were further organized into section-specific subthemes defined qualitatively based on the dominant methodological focus of papers, such as architectural families, metric types, modality interactions, or application domains.
- **Comparative Framework:** Studies were grouped by methodological similarity and compared within internally coherent clusters. This clustering allowed for the identification of dominant trends, recurring evaluation patterns, open research challenges, and consensus or divergence in methodological approaches.
- **Results Presentation:** Synthesis results are presented through a combination of thematic tables (summarizing key attributes, contributions, and evaluation contexts) and detailed narrative sections that integrate findings across papers within each cluster.
- **Evidence Level:** The review does not employ formal tools for assessing risk of bias or certainty of evidence (such as GRADE), as these are typically applied in meta-analyses or highly structured evidence syntheses. Instead, a structured per-study quality appraisal was conducted within each section workflow using a pre-defined set of response categories: *clear, adequate, strong, some, weak or unclear, robust, moderate concerns, and high concerns*. Each retained study was appraised independently by the responsible section lead along dimensions of methodological transparency, evaluation rigour, reproducibility, and contribution clarity. Section narratives acknowledge evidence-base limitations qualitatively, explicitly noting where findings are supported by multiple independent studies versus single isolated reports.

In the final stage, the retained papers within each major thematic category were further organized into section-specific subthemes. These subthemes were defined qualitatively based on the dominant methodological focus of the papers, such as architectural families, metric types, modality interactions, or application domains. The synthesis therefore followed a structured narrative strategy: studies were grouped by methodological similarity, compared within internally coherent clusters, and then interpreted at the section level to identify dominant trends, recurring evaluation patterns, and open research challenges. Narrative synthesis was chosen because the included studies were highly heterogeneous in design, task definition, evaluation protocol, and reported outputs, making formal statistical pooling inappropriate for the present review. This allowed each major section to be structured around coherent internal clusters rather than a simple chronological list of

publications. The overall identification, assignment, screening, and synthesis workflow adopted in this review is summarized in Figure 3.

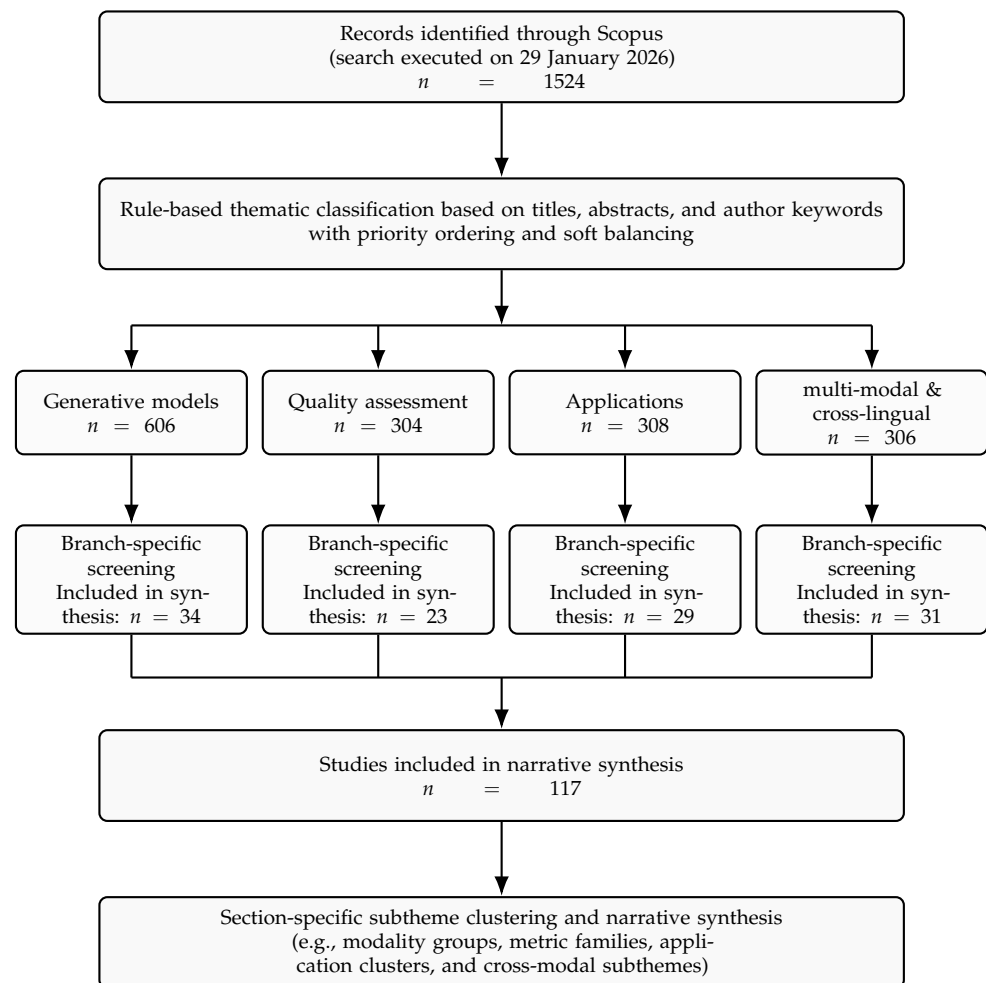


Figure 3. PRISMA-aligned overview of record identification, thematic classification, branch-specific screening, and inclusion in narrative synthesis.

Registration and Protocol Statement: This review was not prospectively registered in PROSPERO, PubMed Central, or any other public registry. A formal review protocol was not prepared prior to the execution of the study. This is explicitly disclosed in order to maintain transparency regarding the review’s retrospective documentation and to acknowledge the constraints this places on claims regarding pre-registered methodological decisions. However, the methodology described above was implemented consistently across all sections, and the search strategy, thematic classification rules, branch-specific screening criteria, and synthesis approach were applied uniformly to all 1524 retrieved records.

2.4. Tools, Software, and Data Availability

Workflow and Documentation Tools: Study selection, data characterization, and synthesis assembly were conducted using a structured spreadsheet-based screening and appraisal workbench that documents section-level appraisals, thematic assignment rules, screening decision rationale, and extraction context for all 117 retained publications. This internal record functions as a machine-readable audit trail for the review methodology.

Data Availability: The final retained corpus of 117 publications is presented in full within the narrative sections and Supplementary Tables S1 and S2 of this manuscript. Study characteristics, model families, evaluation contexts, and thematic classifications are documented throughout the corresponding sections. For researchers seeking additional

methodological detail beyond the present narrative, section-level appraisal notes, excluded-studies summaries, and decision records are available from the corresponding author upon reasonable request, subject to source-database licensing and submission constraints.

Supplementary Note on Excluded Studies (PRISMA Item 16b): The four thematic branches applied branch-specific exclusion criteria to isolate studies directly aligned with each section’s analytical scope. Cross-branch exclusion patterns common to all sections included studies focused on image restoration, denoising, super-resolution, image-to-image translation, medical imaging, remote sensing, deepfake detection, forensics, watermarking, copyright and legal analysis, and general computer-vision or image-processing pipelines without an explicit generative focus. Branch-specific exclusion patterns were as follows.

Generative Models and Architectures (Section 3; 606 assigned → 34 retained): Branch-level screening excluded works primarily focused on deepfake detection or Generative Model fingerprinting, watermarking and content provenance, copyright or legal-responsibility analysis, and low-level optimization studies whose primary contribution lacked a substantive architectural advance in visual synthesis. Representative excluded examples include adversarial-detection papers that use generative networks as a substrate rather than proposing synthesis architectures and benchmark or evaluation studies with no novel generative component.

Quality Assessment and Performance Metrics (Section 4; 304 assigned → 30 candidates → 23 retained): Stage 1 excluded deepfake detection, forensic analysis, watermarking, optimization of the generative-model without a quality assessment objective, and general vision tasks without explicit AIGI evaluation goals, reducing the pool to 30 candidates. Stage 2 removed papers that addressed image quality in general natural-image settings rather than AI-generated content specifically.

Applications (Section 5; 308 assigned → 103 candidates → 29 retained): Branch-level screening excluded articles in which image generation was incidental or only marginally analyzed rather than serving as the central mechanism, studies limited to system descriptions without substantive experimental or evaluative evidence, and works whose primary focus was on a domain (e.g., medical imaging, remote sensing, industrial inspection) already excluded by the global query. A domain-oriented refinement further reduced the candidate pool from 103 to 29 by retaining only studies that reported substantive application evidence in a reasonably broad creative, design, or professional domain.

Multi-modal and Cross-Lingual Aspects (Section 6; 306 assigned → 74 candidates → 31 retained): A two-step branch-specific refinement was applied. Step 1 removed studies focused purely on classification, sentiment analysis, visual question answering without a generative component, robotics, or perception-only multi-modal tasks, yielding 74 candidates. Step 2 retained only papers explicitly addressing multi-modal generation, cross-modal alignment, multi-lingual or cross-lingual transfer, or unified multi-modal foundation architectures relevant to AI-generated content, yielding 31 retained studies. A structured exclusion record covering 188 excluded studies at the earlier branch-screening stage survives and is available upon request; one explicit borderline record is also documented.

The complete excluded-studies record for the multi-modal branch, including short per-record exclusion reasons, is available upon request from the corresponding author and can be included in Supplementary Materials if required by the journal. Per-section evidence-quality summaries based on structured per-study appraisals for the retained corpora of Sections 4–6 are reported in Supplementary Table S2.

3. Generative Models and Architectures for Image Synthesis

This section reviews the main architectural directions that have shaped recent progress in AI-driven image synthesis. The discussion covers representative advances in text-to-

image, face, panoramic, video, and 3D generation, with emphasis on the modeling choices that improved controllability, realism, semantic consistency, and cross-modal conditioning.

Within the global review workflow described in Section 2, the branch of the generative-model contained 606 records after thematic assignment. The section-specific refinement applied here focused exclusively on studies whose primary contribution concerned the design of generative pipelines or architectural mechanisms for visual synthesis across images, faces, panoramas, video, and 3D content. Papers focusing mainly on deepfake detection, watermarking, copyright or legal issues, or low-level optimization without a substantive architectural contribution were excluded during this branch-specific screening. This refinement yielded a final corpus of 34 representative studies for detailed narrative synthesis. To provide sufficient architectural context for interpreting that retained corpus, this section also references eight seminal, field-defining models whose principal contributions are foundational to the modalities under review: Latent Diffusion Models [1], DALL-E 2 [2], Imagen [3], ControlNet [4], DreamBooth [5], SDXL [6], IP-Adapter [7], and DreamFusion [8]. These eight references are discussed as architectural context and are not counted within the 34-study branch corpus or the 117-study PRISMA synthesis total.

The retained generative-model corpus was categorized into five primary groups: (i) text-to-image synthesis, focusing on latent diffusion and unified autoregressive transformers; (ii) facial image generation, emphasizing identity preservation and gaze control; (iii) panoramic synthesis, addressing 360-degree spatial continuity; (iv) video generation, focusing on spatiotemporal transformers and world models; and (v) 3D scene and object synthesis, which encompasses both text-to-3D and 2D-to-3D reconstruction. Each retained article was evaluated based on its methodological core, specific generative primitives (e.g., GANs, NeRFs, or diffusion transformers), and key architectural contributions. Table 1 summarizes both the retained studies and the foundational contextual references discussed in this section, while the subsequent narrative synthesis at the subsection level compares the findings of these internally coherent groups.

Table 1. Summary of representative works on generative models and architectures for image synthesis, including foundational contextual references (Part I) & (Part II).

Part I				
Reference	Year	Modalities/Data	Methodological Core	Key Contribution
<i>Text-to-Image Synthesis</i>				
Rombach et al. [1]	2022	Text→Image	Latent Diffusion Models (LDM)	Performs diffusion in latent space, enabling efficient high-resolution synthesis; architectural foundation of Stable Diffusion.
Ramesh et al. [2]	2022	Text→Image	CLIP-guided hierarchical diffusion (DALL-E 2)	Establishes CLIP latent priors as a conditioning bridge for semantic text-to-image generation.
Saharia et al. [3]	2022	Text→Image	Cascaded diffusion with T5-XXL language model (Imagen)	Demonstrates language model scale as the dominant driver of prompt fidelity in cascaded diffusion.
Zhang et al. [4]	2023	Text + edge/depth/pose→Image	Zero-convolution adapter conditioning (ControlNet)	Enables spatially-aligned auxiliary conditioning on frozen diffusion backbones without degrading generative quality.
Ruiz et al. [5]	2023	Subject images + text→Image	Subject-driven full fine-tuning with prior preservation (DreamBooth)	Achieves identity-preserving personalization from 3–5 reference images via rare-token binding and prior preservation loss.
Podell et al. [6]	2023	Text→Image	Dual-encoder scaled latent diffusion (SDXL)	Improves realism and compositional accuracy via a larger U-Net, dual text encoders, and a two-stage base-refiner pipeline.
Ye et al. [7]	2023	Image + text→Image	Decoupled cross-attention image-prompt adapter (IP-Adapter)	Adds lightweight image-prompt conditioning to frozen diffusion models, composable with ControlNet and LoRA adapters.
Ding et al. [9]	2021	Text→Image	Transformer-based text-to-image generation	Early transformer success in large-scale T2I.

Table 1. Cont.

Part I				
Reference	Year	Modalities/Data	Methodological Core	Key Contribution
Meta AI [10]	2024	Text + Image tokens	Unified Autoregressive Transformer	Generates images and text interchangeably using discrete tokens.
Esser et al. [11]	2024	Text→Image	Rectified Flow Transformers (MM-DiT)	Optimizes denoising trajectories for high-fidelity synthesis.
Black Forest Labs [12]	2025	Text→Image	Flow Matching in Latent Space	Strong prompt adherence and text rendering in open-weight image synthesis.
Lian et al. [13]	2024	Text + layout boxes	LLM-grounded layout guidance	Enables complex spatial reasoning and numeracy in T2I.
Kou et al. [14]	2024	Story prompts→Image	Zero-shot interactive story visualization	Automates character interaction depiction.
Sheng et al. [15]	2024	Text→Image	GPT-enriched text-to-image GAN fusion	Enhances image quality with linguistic enrichment.
Li et al. [16]	2025	Text + layout + identity	Diffusion with layout control for multi-subject scenes	Facilitates personalized multi-subject image generation.
Liu et al. [17]	2025	Text→Image prompts	Residue removal in text-image diffusion	Improves generation by removing concept residue patterns.
Praveen et al. [18]	2025	Text→Image; multi-model API	Multi-modal generative architectures survey	Presents integrated solutions for heterogeneous inputs.
Kumar et al. [19]	2025	Text→Image review	GAN vs transformer comparison in text-image	Benchmarks generative paradigms for T2I.
<i>Facial Image Generation</i>				
Wang et al. [20]	2024	Text + sketch→Face	Sketch-conditioned face diffusion	Incorporates gaze and head pose control via 3D facial sketches.
Igmoullan et al. [21]	2025	Text + sketch→Face	GAN + BERT Transformer fusion	Integrates multi-modal sketch and text cues for refined synthesis.
Wang et al. [22]	2024	Face embedding + text	InstantID (Identity-Preserving Diffusion)	Enables zero-shot identity consistency without model fine-tuning.
<i>Panoramic Synthesis</i>				
Ye et al. [23]	2024	Text→Panorama	Spherical Epipolar Diffusion	Enforces multi-view geometric consistency for ERP synthesis.
Li et al. [24]	2023	Room descriptions→360° scenes	Recursive Outpainting	Generates diverse VLN environments via 360-degree expansion.
Zhang et al. [25]	2024	Text→Panorama	Dual-branch diffusion (PanFusion)	Adapts Stable Diffusion to text-conditioned 360-degree panorama synthesis with improved geometric consistency.
Part II				
Reference	Year	Modalities/Data	Methodological Core	Key Contribution
<i>Video Generation</i>				
Hong et al. [26]	2023	Text→Video	Transformer pretraining for video generation	Large-scale pretraining facilitates T2V quality.
Villegas et al. [27]	2023	Prompt sequences→Video	Variable length video from open text	Generates videos of arbitrary length from text.
Ma et al. [28]	2024	image–text pairs; text→Video	Video Diffusion Transformer	High-speed, realistic motion synthesis from image–text pairs.
Saini et al. [29]	2024	Text→Video	Spatiotemporal Transformers	Scalable 3D-patch architecture for motion coherence.
Bar-Tal et al. [30]	2024	Text→Video	Space-Time U-Net architecture	Ensures temporal consistency via global video denoising.
OpenAI [31]	2024	Text→Video	Diffusion Transformers as World Simulators	Emergent physical simulation and 3D consistency.
Yang et al. [32]	2025	Text + video	Expert Transformer blocks	Prevents semantic interference between text/video modalities.
Kondratyuk et al. [33]	2024	Text + Image + video + audio	LLM conditioned on visual prompts	Uses large language models for generic video generation.
Li et al. [34]	2025	Text + tracklets→Video	Tracklet conditioned video diffusion	Improves temporal consistency via tracklets.
Su et al. [35]	2024	Text→Video models	Pruning strategy for faster generation	Reduces compute while retaining synthesis quality.
Bardes et al. [36]	2025	Video data	Joint-Embedding Predictive Architecture	Advances non-generative physical world modeling.

Table 1. Cont.

Part II				
Reference	Year	Modalities/Data	Methodological Core	Key Contribution
<i>3D Scene and Object Synthesis</i>				
Poole et al. [8]	2023	Text→3D	Score Distillation Sampling on NeRF (DreamFusion)	Establishes SDS as the paradigm for text-to-3D generation without 3D training data; foundational starting point for optimization-based 3D synthesis.
Zhang et al. [37]	2024	Text→3D scene	Tri-planar Radiance Field Optimization	Supports 6-DOF navigation in consistent indoor/outdoor scenes.
Feng et al. [38]	2023	Text→3D layout	LLM-based In-context Layout Planning	Maps complex spatial/numerical relations into 3D indoor scenes.
Sargent et al. [39]	2023	2D images→3D-aware generation	VQ-VAE + Conditional NeRF decoder	Enables stable 3D-aware generation from large 2D image datasets.
Hong et al. [40]	2024	Single image→3D	Transformer-based Feed-forward NeRF (LRM)	Achieves instant (sub-5s) 3D reconstruction from a single image.
He et al. [41]	2025	Single image→3D assets	Semantic-aware LRM (S-LRM)	Reconstructs semantically decomposed 3D assets in a single pass.
Shi et al. [42]	2024	Text + multi-view priors→3D	Multi-view Diffusion with 3D Self-Attention	Mitigates the Janus problem via 3D-aware generative priors.

3.1. Section-Specific Selection and Thematic Grouping

The following subsections provide a detailed synthesis of the findings of the selected articles, organized according to the identified thematic categories.

3.2. Text-to-Image Generation

Text-to-image generation has emerged as one of the most advanced and widely deployed applications of modern generative modeling. The field has undergone two major paradigm shifts in a relatively short period of time. The first was the transition from generative adversarial networks (GANs) to diffusion-based generative models, and the second was the shift from performing denoising directly in pixel space to performing it in a learned latent representation space. A precise understanding of the factors driving these transitions, as well as the associated trade-offs in sample quality, training stability, data efficiency, and computational cost, is a necessary prerequisite for a nuanced interpretation of the methodological contributions reviewed in the following sections.

GANs delivered fast, high-fidelity image synthesis through adversarial training but were notoriously unstable and prone to mode collapse, limiting their ability to model the full data distribution [43]. Diffusion models largely mitigated these limitations through iterative denoising, trading inference speed for improved sample diversity, and training stability [44]. Latent Diffusion Models (LDMs) [1] subsequently addressed much of the computational burden by performing diffusion in a low-dimensional latent space learned by an auto encoder, thus making large-scale deployment and open-weight release practical. This shift to latent-space diffusion is arguably the most consequential architectural development of the diffusion era: it reduced computational requirements by an order of magnitude while establishing cross-attention as the canonical mechanism for conditioning generation on text, style, layout, and identity information. Since then, three principal axes of progress have emerged: conditioning richness, spatial controllability, and personalization. All are grounded in the latent-diffusion paradigm and serve as the organizing backbone of the literature reviewed in this survey.

In 2021, Ding et al. [9] introduced CogView, a 4-billion-parameter Transformer that uses a vector quantized variational auto encoder tokenizer (VQ-VAE). To stabilize the training of this large model, they proposed Precision Bottleneck Relaxation (PB-Relax) and Sandwich LayerNorm (Sandwich-LN), two techniques designed to mitigate numerical overflow and underflow issues. Beyond image synthesis, the model supports fine-tuning

for a variety of downstream tasks, including super-resolution, style transfer, image captioning, and text–image re-ranking, highlighting its versatility and applicability to real-world domains such as industrial fashion design. Furthermore, the authors introduced Caption Loss (CapLoss), a novel metric for evaluating text–image alignment, and proposed strategies to address fairness concerns in generative modeling. Collectively, these contributions advance controllable image generation and deepen understanding of cross-modal representation learning.

The dominant paradigm in contemporary text-to-image synthesis was established by Rombach et al. [1], who introduced Latent Diffusion Models (LDMs). Rather than performing the denoising process in high-dimensional pixel space, LDMs compress images into a low-dimensional perceptual latent space via a pretrained auto encoder and apply the diffusion process there, dramatically reducing computational cost without sacrificing perceptual quality. Cross-attention layers inserted into the U-Net backbone allow arbitrary conditioning signals, including text, class labels, and semantic maps, to guide synthesis. This architectural decision became the foundation for Stable Diffusion and essentially all subsequent large-scale open-weight text-to-image systems, making LDMs the single most influential architectural contribution in the diffusion era of image synthesis.

Building on the success of CLIP for cross-modal alignment, Ramesh et al. [2] introduced DALL-E 2, which decomposes text-to-image generation into a prior network that maps CLIP text embeddings to CLIP image embeddings, followed by a diffusion decoder that renders images from those image embeddings. This hierarchical design enables strong semantic alignment and compositionality, as the CLIP latent space encodes high-level visual concepts independently of low-level pixel statistics. DALL-E 2 demonstrated the use of rich joint text–image embedding spaces as an intermediate conditioning bridge, rather than conditioning the diffusion model on raw text directly. This substantially improved prompt fidelity and image coherence, and the model produced photorealistic output at a quality level that established it as a landmark in CLIP-guided text-to-image generation.

Concurrently, Saharia et al. [3] proposed Imagen, which conditions a cascaded diffusion pipeline not on CLIP embeddings but on representations from large pretrained language models, specifically T5-XXL. This demonstrated that the richness of a text encoder matters more for semantic fidelity than the scale of the image diffusion model alone. Imagen employs a base 64×64 diffusion model followed by two super-resolution diffusion stages, collectively enabling high-fidelity 1024×1024 synthesis from text prompts. The work introduced the DrawBench benchmark for evaluating compositional text–image alignment and showed that language model scale and classifier-free guidance strength are the two most impactful variables for perceptual quality and semantic accuracy, establishing Imagen as an important milestone in the scaling-law-based development of text-to-image systems.

Podell et al. [6] introduced SDXL, which scales the Stable Diffusion architecture along several dimensions simultaneously: a significantly larger U-Net backbone with a three-times increase in attention blocks, an ensemble of two text encoders (OpenCLIP ViT-bigG and CLIP ViT-L), and a two-stage pipeline in which a base model generates a 1024×1024 latent that is subsequently refined by a latent-space refiner model. SDXL also introduces conditioning on original image resolution and crop coordinates to prevent the model from learning resolution-dependent biases that arise when training images rescaled to a common size. The result is substantially improved realism, compositional accuracy, and adherence to complex prompts compared to earlier Stable Diffusion versions, making SDXL the de facto open-weight baseline for high-resolution single-image generation at the time of its release.

Although diffusion models dominate the current landscape, a significant parallel evolution has occurred in unified multi-modal architectures. Meta AI [10] publicly released Chameleon, a family of mixed-modal foundation models that utilize an early-fusion token-

based approach. Unlike diffusion-based systems, Chameleon treats images and text as discrete tokens within a single transformer, allowing seamless interleaved generation. This unified autoregressive paradigm enables the model to reason across modalities without separate encoders or decoders, representing an important step toward general-purpose multi-modal intelligence.

Recent advances have shifted the focus to architectural scaling and optimization of noise trajectories. Esser et al. [11] introduced the Scaling Rectified Flow Transformer (SD3), which replaced traditional U-Net structures with a multi-modal Diffusion Transformer (MM-DiT). Using Rectified Flow, the model learns a straight-line trajectory between noise and data, significantly improving sampling efficiency and image fidelity. Building upon this trajectory-based paradigm, Black Forest Labs [12] released FLUX.1 Kontext, which utilizes Flow Matching in latent space to improve prompt adherence and text rendering while maintaining strong open-weight image synthesis performance.

A complementary line of work has focused on adding fine-grained spatial controllability to pretrained diffusion models without degrading their generative capacity. Zhang et al. [4] introduced ControlNet, which attaches a trainable copy of the encoding layers of a frozen diffusion backbone to accept auxiliary spatial conditioning signals, including edge maps, depth maps, human pose skeletons, and segmentation masks. This is implemented through zero-initialized convolution layers (zero convolutions), whose design ensures that at the start of training the ControlNet branch contributes zero signal to the frozen backbone, preventing the corruption of learned generative priors during adaptation. ControlNet demonstrated that spatially aligned conditioning can be added to large pretrained diffusion models in a modular, composable manner, establishing the adapter-based controllability paradigm that has since been widely adopted and extended.

For subject-level personalization, Ruiz et al. [5] proposed DreamBooth, which fine-tunes an entire text-to-image diffusion model on a small set of three to five images of a specific subject, binding the subject's visual identity to a rare token in the text embedding space. A key contribution is the introduction of a class-specific prior preservation loss that prevents language drift, i.e., the tendency of full fine-tuning to overfit to the few reference images and lose the ability to generate semantically related but distinct instances of the same class. DreamBooth demonstrated that strong subject fidelity across diverse poses, styles, and contexts can be achieved through full-model adaptation rather than test-time optimization, making it a key milestone in identity-preserving personalization that has since inspired numerous downstream methods for characters, objects, and styles.

Ye et al. [7] proposed IP-Adapter as a lightweight alternative to full fine-tuning for image-prompt conditioning. The method introduces a decoupled cross-attention mechanism in which a separate set of cross-attention layers handles image features extracted by a pretrained CLIP image encoder, while the original text cross-attention layers remain unchanged; the two pathways are then summed to produce the final conditioning. Because only the added cross-attention parameters are trained, IP-Adapter is efficient and composable. It can be combined with ControlNet or LoRA adapters without conflict, and it can be generalized across different fine-tuned versions of Stable Diffusion without retraining. IP-Adapter demonstrated that image-prompt conditioning rivaling full fine-tuning can be achieved with a small fraction of the parameters and training compute, and it has since become a foundational building block for style transfer, face-driven generation, and multi-modal conditioning pipelines.

To address the persistent challenge of spatial reasoning in diffusion models, Lian et al. [13] proposed an LLM-based diffusion framework. This approach leverages the linguistic reasoning capabilities of Large Language Models to first generate a structured layout (bounding boxes) from a text prompt, which then serves as a grounded guide for the

diffusion process. This two-stage method effectively mitigates “prompt neglect” and allows for the precise positioning of multiple subjects within a single scene, a task that traditional end-to-end models often struggle to perform accurately.

Kou et al. [14] examined automated multi-character zero-shot story visualization, focusing on how to generate portraits that remain both distinctive and mutually compatible within a shared narrative. Their LeMon framework combines LLM-based character initialization with a graph-based text-to-image diffusion model that explicitly represents character interactions. This combination improves the consistency of multi-character story imagery while reducing the need for manual setup.

Sheng et al. [15] introduced ISF-GAN, a GAN-based text-to-image framework built around text enrichment and cross-modal fusion. The method first expands sparse textual descriptions with a GPT-based language model and then filters and fuses the enriched descriptions through cross-modal attention over multi-scale image features. The main contribution is to recover missing semantic information prior to generation, thus improving image–text consistency.

Li et al. [16] proposed LCP-Diffusion, a tuning-free framework for layout-controllable personalized image synthesis. The method combines identity preservation with explicit spatial control by extracting complementary dynamic and static subject features and then steering cross-attention toward user-specified layout regions. This design improves multi-subject placement and identity fidelity while reducing copy-and-paste artifacts.

Liu et al. [17] proposed Corer, a concept-erasure framework designed to prevent diffusion models from regenerating removed concepts under semantically related prompts. Rather than suppressing only the target concept, the method also handles associated concepts while regularizing unrelated regions of the latent space to preserve generation specificity. In this way, Corer reduces the concept residue without broadly degrading the fidelity of unrelated content.

Praveen et al. [18] proposed AI ImageGen, a user-facing framework for creating images with multiple AI models. By integrating several models through the Hugging Face API, the system allows users to generate images from custom text prompts and to manage the resulting outputs through standard viewing, downloading, and deletion functions. Its main contribution lies in exposing model choice as part of the user workflow, thereby promoting output diversity and improving accessibility for non-technical users interested in AI-assisted image creation.

Kumar et al. [19] compared GAN- and transformer-based models for text-to-image generation, evaluating how well each converts text into realistic images. GANs use a generator–discriminator framework and improve image fidelity via multi-stage generation and attention, but struggle with complex datasets. Transformers, originating in NLP, better model dependencies, improving feature extraction from text. Emerging hybrids combine both to enhance both stability and image quality. The study analyzes architectures, training, and output quality to identify optimal techniques for multi-modal generative modeling.

Among these contributions, three design choices emerge as the most influential. First, the choice of conditioning representation: CLIP image embeddings (DALL-E 2 [2]), large language model text embeddings (Imagen [3]) or dual-encoder ensembles (SDXL [6]). These choices determine the limit on prompt fidelity and compositional accuracy. Evidence consistently favors the language model scale over the image model scale as the primary driver of semantic alignment. Second, the decision between full fine-tuning and lightweight adaptation governs the controllability–generality trade-off. DreamBooth [5] achieves maximum subject fidelity through full-model adaptation at the cost of language drift, while ControlNet [4] and IP-Adapter [7] preserve the generative prior by routing auxiliary signals through frozen-backbone adapters, enabling modular composition but

sacrificing the depth of personalization achievable through full fine-tuning. Third, the shift from U-Net to transformer backbones (SD3 [11], FLUX [12]) reflects a broader architectural bet on attention-based scaling: transformers handle long-range spatial dependencies and multi-modal token mixing more naturally than convolutional hierarchies, but they increase memory footprint and require larger training budgets to reach the same perceptual quality. The remaining open problems are compositional multi-object accuracy [13], concept interference after erasure [17], and unified multi-subject identity control [16]. These all point to the same root limitation: text-to-image diffusion models learn correlations in joint embedding space rather than explicit compositional structure, and closing this gap will likely require richer intermediate representations than current cross-attention affords.

3.3. Text-to-Face Generation

One specific application of text-to-image algorithms is the generation of faces. For example, Wang et al. [20] present a gaze-controllable face generation algorithm that takes as input textual descriptions of human gaze and head behavior and synthesizes corresponding facial images. The model first constructs a text-of-gaze dataset comprising over 90k textual descriptions that densely cover the distribution of gaze directions and head poses. The proposed framework consists of a sketch-conditioned face diffusion module and a model-based sketch diffusion module. The facial sketch is defined using facial landmarks and an eye segmentation map, providing a structured and fine-grained representation that serves as a strong prior for subsequent face synthesis. The face diffusion module generates high-quality face images from this facial sketch, while the sketch diffusion module leverages a 3D face model to infer a facial sketch directly from the textual description.

Igmoullan et al. [21] propose a model to generate human face images from text descriptions and sketches by combining transformer-based language models with Generative Adversarial Networks (GANs). Their approach uses a StackGAN framework conditioned on embeddings derived from textual captions and visual sketch inputs. Text semantics are encoded with a BERT transformer, while sketch features are extracted via a pre-trained VGG16 model. The multi-modal embeddings are fused and fed into a two-stage GAN to progressively synthesize high-resolution facial images. Adversarial, Kullback–Leibler divergence, and triplet-margin losses are combined to promote realism, diversity, and identity preservation during training.

However, the field has recently shifted toward tuning-free identity preservation, moving away from rigid landmarks toward semantic identity embeddings. Wang et al. [22] introduce InstantID, which uses a decoupled cross-attention mechanism to integrate face embeddings from pre-trained recognition models such as InsightFace into a latent diffusion pipeline. Unlike earlier GAN-based methods, this approach allows the generation of a specific individual in any pose or style defined by a text prompt without requiring model fine-tuning. This represents a broader shift toward zero-shot identity consistency, where the model maintains high-frequency facial details across diverse textual contexts while preserving the underlying generative prior of large-scale diffusion models.

3.4. Text-to-Panoramic Image Generation

Another modality requiring specialized spatial priors is text-guided panoramic image generation, a task crucial for immersive media and virtual reality. The primary challenge lies in maintaining global boundary consistency and mitigating the geometric distortion inherent in Equirectangular Projections (ERP). Traditional approaches, such as fixed-camera single-view generation or naive multi-perspective stitching, frequently fail to maintain 3D awareness or multi-view consistency.

Ye et al. [23] addressed these limitations by introducing DiffPano, a framework that fine-tunes Stable Diffusion using Low-Rank Adaptation (LoRA) in panoramic datasets. By incorporating a spherical epipolar-aware multi-view diffusion model, DiffPano enforces geometric constraints across generated views, ensuring scalable and coherent panoramic scenes. Similarly, Li et al. [24] proposed PanoGen, which utilizes recursive outpainting to create 360-degree environments for vision–language Navigation (VLN). By conditioning on room descriptions from Matterport3D, PanoGen generates diverse semantically consistent layouts that significantly enhance the robustness of VLN agents in unseen environments.

Although recursive methods like PanoGen provide diversity, they often accumulate semantic drift at the stitching boundary. More recent work has therefore moved toward latent architectures adapted specifically for 360-degree panorama synthesis. Zhang et al. [25] introduced PanFusion, a dual-branch diffusion framework that adapts Stable Diffusion to text-conditioned panoramic image generation. By combining perspective and equirect-angular branches, the method improves visual fidelity while better preserving geometric consistency across the panoramic field of view. This direction illustrates how text-to-image diffusion can be specialized for 360-degree content without relying only on recursive outpainting or post hoc stitching.

3.5. Text-to-Video Generation

Text-to-video generation did not emerge as a clean successor to image synthesis; it required a rethinking of both the generative objective and the architectural unit of computation. The core difficulty is that video adds a temporal axis along which motion must be physically plausible, semantically consistent, and controllable—constraints that image diffusion models satisfy trivially by construction, but that video models must learn explicitly. Three paradigm shifts characterize the trajectory of the corpus reviewed here. The first was the move from frame-level generation and interpolation (CogVideo [26], Phenaki [27]) toward joint spatiotemporal denoising, which replaced the compounding error of recursive keyframe interpolation with a single coherent generation pass over the full clip (Lumiere [30], Latte [28]). The second was the reframing of video generation as world simulation: treating video not as a sequence of images, but as a compressed representation of physical dynamics, with spacetime patches as the atomic unit (Sora [31], Snap Video [29]). The third, still nascent, is the move from purely generative objectives to predictive world models that learn physical structure without pixel-level reconstruction (V-JEPA 2 [36]). The key architectural trade-off throughout is between temporal resolution and training tractability: architectures that process entire clips jointly achieve stronger coherence but scale poorly, while those that operate frame-by-frame or through keyframe interpolation are efficient but accumulate drift.

Hong et al. [26] introduce CogVideo, a 9-billion-parameter Transformer-based text-to-video generation model developed leveraging a pretrained text-to-image backbone, CogView2. Instead of training from scratch, CogVideo “inherits” knowledge from a pretrained text-to-image model, CogView2. This significantly reduces training costs and provides a strong foundation for spatial semantics. The model uses a multi-frame-rate training strategy. It first generates a few sparse key frames and then recursively interpolates them to create a smooth, high-frame-rate video. It can generate 4 s video clips (32 frames) at approximately 8 frames per second.

Villegas et al. [27] introduce Phenaki, a model for text-conditioned video synthesis from sequences of prompts. Its key design choice is a compact token representation of video with causal temporal attention, which makes variable-length generation more tractable. By combining image–text and video–text training data, the framework extends video duration and prompt flexibility beyond what earlier methods could reliably support.

The efficiency and accessibility of high-fidelity motion synthesis is further advanced by Ma et al. [28], who introduce Latte, a diffusion-transformer-based model that enables scalable video generation. Their work illustrates a broader shift toward architectures that preserve the structural consistency of an initial frame while synthesizing continuous motion, thereby narrowing the gap between static text-to-image outputs and temporally coherent video generation.

Significant architectural changes occurred in 2024 with the introduction of spatiotemporal transformers. Snap Video [29] treats video as 3D patches to improve scaling and motion coherence, while Lumiere [30] generates an entire clip jointly rather than through hierarchical keyframe interpolation. Sora [31] further popularized the framed video generation as a world-simulation task based on spacetime patches, highlighting the emergent physical and geometric consistency. CogVideoX [32] then refined this transformer line by separating text and video parameters to reduce semantic interference.

Kondratyuk et al. [33] introduce VideoPoet, a model designed to synthesize high-quality videos from a broad spectrum of conditioning signals. VideoPoet adopts a decoder-only Transformer architecture capable of processing multi-modal inputs, including images, videos, text, and audio. Its training protocol follows the paradigm of Large Language Models (LLMs) and comprises two principal stages: large-scale pretraining and subsequent task-specific adaptation. During pretraining, VideoPoet optimizes a mixture of multi-modal generative objectives within an autoregressive Transformer framework. The resulting pre-trained LLM backbone serves as a general-purpose foundation that can be further adapted to diverse video generation tasks. Empirical results indicate that the model achieves state-of-the-art performance in zero-shot video generation, with particular strengths in producing high-fidelity motion dynamics.

Li et al. [34] introduce TrackDiffusion, a video generation framework that enables fine-grained, trajectory-conditioned motion control via diffusion models. This approach supports precise manipulation of object trajectories and interactions, thereby addressing common issues related to scale inconsistencies and disruptions in motion continuity. A key component of TrackDiffusion is the instance enhancer, which explicitly enforces inter-frame consistency across multiple objects, an aspect largely neglected in prior work. The authors also show that the video sequences synthesized by TrackDiffusion can serve as effective training data for visual perception models. More broadly, the work demonstrates that video diffusion models conditioned on tracklets can improve the performance of downstream object tracking systems.

Su et al. [35] investigate the inference mechanisms of two main T2V architectures based on transformers and diffusion models. Their analysis reveals substantial redundancy in the temporal attention modules of both architectures, which are typically employed to capture temporal dependencies across video frames. Building on these findings, the authors introduce a model-agnostic and training-free pruning framework, termed F3-Pruning, designed to remove redundant temporal attention weights. Concretely, when aggregated temporal attention scores fall below a predefined percentile threshold, the associated attention weights are pruned. Comprehensive experiments conducted on three benchmark datasets, using the canonical transformer-based model CogVideo and the representative diffusion-based model Tune-A-Video, demonstrate that F3-Pruning achieves significant inference-time acceleration while maintaining output quality and exhibiting strong generalization across different model families.

Finally, moving beyond purely generative objectives, Bardes et al. [36] introduce V-JEPA 2. This work utilizes a Joint-Embedding Predictive Architecture to learn a latent world model by predicting missing parts of a video. By avoiding pixel-level reconstruction, V-JEPA 2 provides a more robust representation of physical interactions and action-

conditioned world physics, bridging the gap between video synthesis and autonomous world understanding.

Together, the video corpus reveals a clear hierarchy of design choices. The most consequential decision is whether to process time as an independent axis (inflated 2D attention, as in CogVideo [26]) or as a fully coupled spatial-temporal tensor (3D patch transformers, as in Sora [31] and Snap Video [29]): the former is more data-efficient and transfers better from image pretraining, but the latter is necessary for emergent physical consistency across many frames. The second critical choice is the scope of training data: models trained on video-text pairs alone (CogVideo [26], CogVideoX [32]) learn strong semantic grounding but are limited by the scale of the video dataset, while models trained on mixed image-text and video-text corpora (VideoPoet [33]) achieve broader coverage at the cost of weaker temporal specialization. Third, the treatment of motion controllability separates the corpus into two camps: free-form motion learned implicitly from scale (Sora, Lumiere) versus explicit conditioning on structured trajectories or objects (TrackDiffusion [34]); the former produces more cinematic variety but is hard to control, while the latter enables reproducible, fine-grained motion manipulation at the cost of requiring structured inputs at inference time. The efficiency findings of Su et al. [35] further reveal that temporal attention modules are substantially over-parameterized across both architectural families, suggesting that the capacity dedicated to temporal modeling in current models exceeds what is actually exploited, which is a finding that constrains how far scaling alone can be expected to resolve remaining coherence failures.

3.6. Three-Dimensional Scene and Object Synthesis

The synthesis of three-dimensional content from multi-modal inputs represents a critical frontier in generative modeling, enabling richer downstream interactions such as novel-view synthesis, 3D editing, and integration into immersive virtual environments. This task is inherently challenging due to the scarcity of large-scale text-3D paired datasets and the requirement for strict spatiotemporal consistency across diverse viewing angles. The foundational paradigm for lifting 2D diffusion priors into 3D was established by Poole et al. [8] (DreamFusion), which introduced Score Distillation Sampling (SDS)—a technique that back-propagates gradients from a frozen 2D diffusion model into a differentiable 3D representation (a NeRF) by treating the diffusion score function as a loss signal. Because SDS requires only a pretrained 2D diffusion model and no 3D supervision whatsoever, DreamFusion enabled text-to-3D generation for the first time at scale without 3D training data, decoupling the problem of learning 3D geometry from the requirement of collecting paired text-3D datasets. The method suffers from the well-known Janus (multi-face) problem arising from the fact that 2D renderings from different viewpoints are scored independently, but it nonetheless established SDS as the dominant optimization objective in the field and motivated a large body of subsequent work on improved 3D representations, multi-view consistency, and faster convergence. Subsequent research in this subsection can be understood largely as addressing the efficiency and consistency limitations of the SDS-based paradigm introduced by DreamFusion.

For example, Zhang et al. [37] introduced 3D SceneDreamer, which uses a tri-planar radiance-field representation to support consistent 6-DOF scene generation. Feng et al. [38] complemented this line with LayoutGPT, where large language models are used for spatial planning and semantic layout generation in indoor scenes. Sargent et al. [39] further stabilized 3D-aware generation on heterogeneous 2D data through a VQ-VAE-based pipeline with a conditional NeRF decoder.

Despite the success of optimization-based methods, they are often computationally expensive and subject to the Janus problem, where independent view optimizations lead to in-

consistent geometry. To address this, recent research has shifted toward large reconstruction models (LRMs) and multi-view diffusion priors that perform direct feed-forward synthesis. Hong et al. [40] introduced the Large Reconstruction Model (LRM), a transformer-based architecture of 500 million parameters that predicts a 3D NeRF from a single input image in approximately five seconds. By training on a massive dataset of one million objects, LRM achieves strong generalization across diverse real-world and generative inputs. Building on this high-speed reconstruction paradigm, He et al. [41] introduced StdGEN, a pipeline designed to generate semantically decomposed 3D characters. At its core, a semantic-aware LRM (S-LRM) jointly reconstructs geometry and appearance alongside distinct semantic components (e.g., body, clothing, hair) in a feed-forward fashion, enabling the recovery of explicit meshes within three minutes.

In addition to these reconstruction models, Shi et al. [42] proposed MVDream, a multi-view diffusion model that integrates 3D self-attention into the diffusion pipeline. By jointly leveraging 2D and 3D training data, MVDream inherits the generalization of 2D models while maintaining the spatial coherence of 3D renderings. This model acts as an implicit, highly generalizable 3D prior that substantially improves the stability of 2D-to-3D lifting approaches and supports few-shot concept learning for personalized 3D generation. Collectively, these advances transition 3D synthesis from slow, per-scene optimization toward a unified, efficient framework for turning textual and visual prompts into consistent 3D assets.

The 3D corpus is organized around a single dominant paradigm shift: from optimization-based distillation to feed-forward reconstruction. The shift was driven by three compounding limitations of SDS-based approaches: (i) per-scene optimization times on the order of hours render them impractical for interactive or large-scale use; (ii) the Janus problem, which is a multi-face geometry arising from view-independent scoring, requires additional multi-view constraints that DreamFusion did not provide; and (iii) the diversity of outputs is bounded by the 2D diffusion prior, not by any explicit 3D structural model. Feed-forward models (LRM [40], StdGEN [41]) resolved points (i) and (iii) by shifting inference from iterative NeRF optimization to a single transformer forward pass, reducing reconstruction time from hours to seconds and enabling training on datasets large enough to learn genuine 3D shape priors. Multi-view diffusion (MVDream [42]) addressed point (ii) by enforcing view consistency within the denoising process itself, using 3D self-attention to couple renderings from different viewpoints during training. The most influential single design choice in the corpus is the transformer-based NeRF decoder in LRM: by treating triplane NeRF coefficients as the prediction target of a large-scale transformer, LRM unified the representation learning and 3D decoding into a single differentiable pipeline, establishing the architectural template for nearly all subsequent feed-forward 3D reconstruction work. The remaining challenge is the compositional scene generation beyond individual objects, which is addressed by LayoutGPT [38] and 3D SceneDreamer [37] through LLM-guided spatial planning and radiance-field representations, respectively, but neither yet achieves the open-vocabulary generality of the best text-to-image models, indicating that the gap between object- and scene-level 3D synthesis remains a structural open problem.

3.7. Architectural Paradigm Shifts and Trade-Offs in Visual Synthesis

The trajectory of generative modeling is characterized by a series of identifiable paradigm shifts, each resolving the limitations of the preceding approach at the cost of introducing new constraints. Understanding these transitions and their inherent trade-offs is essential for interpreting the contributions of the retained corpus.

From GANs to Diffusion Models. Generative Adversarial Networks [43] dominated image synthesis for nearly a decade, offering fast single-pass inference and high-fidelity out-

put, but suffering from training instability, mode collapse, and limited diversity. Denoising diffusion probabilistic models [44] resolved the diversity and stability problems through iterative denoising, but at the cost of significantly slower sampling due to the multi-step reverse process. Latent Diffusion Models [1] partially addressed the efficiency gap by moving the denoising into a compressed latent space, reducing computational cost while preserving perceptual quality. The core trade-off that remains is sampling speed versus output quality: reducing diffusion steps through DDIM-style distillation or flow-matching trajectories [11,12] recovers some of the inference speed of GANs at the cost of reduced sample diversity or increased sensitivity to trajectory linearization artifacts.

Controllability versus Generative Capacity. A persistent tension in the field concerns how much conditioning control can be imposed on a generative model without degrading its learned prior. Zero-convolution adapters (ControlNet [4]) and image-prompt adapters (IP-Adapter [7]) demonstrate that auxiliary conditioning pathways can be grafted onto frozen backbones with minimal disruption to generative quality; however, each adapter specializes to a signal type (spatial structure, reference style, identity) and composing multiple adapters into a coherent unified conditioning interface remains an open problem illustrated by the multi-subject layout work of [16]. Full fine-tuning approaches such as DreamBooth [5] achieve stronger subject fidelity but risk language drift—reducing the backbone’s ability to respond to novel prompts unrelated to the fine-tuned subject. This fidelity versus generality trade-off is also visible in the face generation subsection, where identity-preserving methods [22] succeed in zero-shot settings precisely by anchoring generation to recognition embeddings rather than retraining the diffusion prior.

Autoregressive versus Diffusion Paradigms. The retained text-to-image corpus reflects a broader competition between diffusion-based and autoregressive generation paradigms. Early transformer-based models such as CogView [9] showed that discrete token autoregressors can achieve compelling text–image alignment on large corpora; unified multi-modal architectures such as Chameleon [10] extend this to mixed-modal sequences. However, diffusion transformers [11] currently dominate high-fidelity single-image synthesis because continuous-latent denoising naturally handles perceptual coherence in a way that discrete token prediction must re-learn through scale. The trade-offs between these paradigms concern generation granularity (token-level composition vs. holistic denoising), training data requirements, and suitability for video: world-simulator-framed models such as Sora [31] and Snap Video [29] favor spatiotemporal continuity over token efficiency.

Specialized versus Unified Architectures. Early work focused on task-specific pipelines for images [9], video [26], or 3D content [39]. Contemporary models such as Sora [31] and Chameleon [10] indicate that scaling spatial-temporal transformers and early-fusion architectures yields stronger physical and semantic reasoning by treating all modalities within a shared representation space. This architectural convergence addresses the semantic drift prevalent in earlier recursive outpainting methods [24], but replaces it with new challenges: highly unified models require massive compute and diverse paired training data, limiting accessibility and domain adaptation. The transition from per-scene optimization-based 3D synthesis (DreamFusion [8]) to large feed-forward reconstruction models (LRMs [40]) illustrates the same compute-accessibility generalization triangle that governs the broader field.

Remaining Challenges. Physical grounding continues to be a bottleneck: while models can generate visually convincing motion, they often struggle with complex causal interactions and fine-grained fluid dynamics, a gap currently being addressed by Joint-Embedding Predictive Architectures (V-JEPA) [36]. Consistent identity across diverse modalities—from 2D faces [22] to 3D characters [41]—demands more efficient decoupling of style, structure, and content. Spatial reasoning and compositional accuracy under com-

plex multi-object prompts remain measurably weaker than perceptual realism, as surveyed in the LLM-grounded layout work [13] and the concept-residue analysis of [17]. As diffusion transformers [11] and flow-matching architectures [12] converge towards unified generation, it is likely that the next generation of visual synthesis will move beyond content creation toward physically consistent digital twins and interactive virtual environments.

3.8. Section-Specific Limitations

The generative-architecture corpus in this section remains heterogeneous with respect to tasks (2D, panoramic, video, 3D), reporting granularity, and evaluation conventions, which limits strict one-to-one comparison across studies. Several contributions emphasize architectural novelty and qualitative demonstrations rather than harmonized benchmark protocols, and some field-defining systems are documented primarily as technical reports at the time of review. Consequently, the conclusions of the paper in this section should be interpreted as a directional synthesis of trends rather than as a uniform, protocol-controlled ranking of methods.

4. Quality Assessment and Performance Metrics for AI-Generated Images

This section examines how AI-generated images are evaluated once generation quality is no longer reducible to traditional distortion-based image quality assessment. The focus is on objective metrics, subjective protocols, prompt-aware predictors, and benchmark design choices that jointly capture perceptual realism, semantic faithfulness, and human preference.

4.1. Section-Specific Selection and Thematic Grouping

Within the global review workflow described in Section 2, the quality assessment branch contained 304 records after thematic assignment. A two-stage section-specific screening procedure was then applied to isolate studies directly concerned with the evaluation of AI-generated images.

In Stage 1, a title- and abstract-level screening applied domain-specific constraints. Studies focusing exclusively on deepfake detection, forensic analysis, watermarking, generative model optimization, or general vision tasks without explicit quality assessment objectives were excluded. This step reduced the pool to 30 candidate studies.

In Stage 2, a stricter semantic consistency check retained only papers explicitly addressing perceptual quality evaluation, text–image semantic alignment, benchmarking protocols, distribution-based realism metrics, cross-modal semantic alignment, or human-centred perceptual and preference modeling. This final screening stage yielded 23 core studies. The retained corpus was then organized into methodological groups spanning prompt-aware and blind AIGIQA models, alignment and faithfulness metrics, human preference and reward models, datasets and benchmarks, and broader survey or context papers, as summarized in Table 2. The table presents the retained study characteristics, and the following narrative synthesis compares results, evaluation logic, and recurring limitations across these groups.

The following subsections provide a detailed synthesis of the findings from the selected articles, beginning with an overview of the field and then proceeding through objective metrics, subjective evaluation, learning-based predictors, benchmark datasets, and open challenges.

Table 2. Summary of representative works on quality assessment and performance metrics for AI-generated images Part I and Part II.

Part I				
Reference	Year	Modalities/Data	Methodological Core	Key Contribution
<i>AIGIQA Models (Prompt-Aware/Blind)</i>				
Zhou et al. [45]	2025	Text + Image	Introduces CIA-Net, a cross-modality interactive attention architecture that models bidirectional text–image feature interactions for prompt-aware quality regression.	Strengthens prompt-aware AIGI quality prediction through explicit bidirectional text–image interaction modeling.
Zhou et al. [46]	2024	Image	Proposes AMFF-Net, an adaptive mixed-scale feature fusion network aggregating multi-level visual representations for blind AIGI quality prediction.	Improves blind AIGIQA robustness by capturing multi-scale distortion patterns without relying on textual prompts.
Yuan et al. [47]	2024	Text + Image	Introduces TIER, a dual text–image encoder regression framework combining vision–language embeddings with quality-aware regression heads.	Validates prompt-conditioned regression as an effective paradigm for perceptual AIGI quality assessment.
Zhang et al. [48]	2024	Text + Image	Proposes a cross-modal correlation modeling framework that explicitly captures statistical dependencies between textual semantics and visual features.	Demonstrates performance gains from explicit cross-modal correlation learning in AIGIQA.
Yu et al. [49]	2024	Text + Image	Introduces SF-IQA, a score-fusion architecture integrating multi-layer perceptual feature extraction with fine-tuned vision–language similarity modeling.	Unifies perceptual fidelity and semantic similarity into a single scoring framework for more holistic AIGI evaluation.
Tang et al. [50]	2024	Text + Image	Proposes CLIP-AIGIQA, a CLIP-enhanced regression model with learnable prompt tokens for strengthened prompt-conditioned quality assessment.	Enhances prompt-aware assessment by leveraging pretrained vision–language representations for stronger alignment sensitivity.
Chen et al. [51]	2025	Text + Image	Introduces QMI-Net within a unified subjective-objective framework combining structured human evaluation with deep regression modeling.	Establishes a structured subjective evaluation protocol aligned with objective modeling for reliable AIGI benchmarking.
Aziz et al. [52]	2025	Text + Image	Proposes a unified human-metric evaluation framework integrating perceptual calibration with computational scoring mechanisms.	Integrates perceptual assessment with computational metrics for comprehensive AIGI evaluation.
Sridhar et al. [53]	2025	Text + Image	Proposes a prompt-aware no-reference evaluation framework integrating multi-granularity similarity modeling with semantic constraints.	Highlights limitations of traditional NR metrics and advances prompt-integrated evaluation for generative content.
<i>Alignment and Faithfulness Metrics</i>				
Hessel et al. [54]	2021	Text + Image	Introduces CLIPScore, a reference-free metric computing cosine similarity between CLIP image and text embeddings.	Establishes a reference-free semantic alignment metric widely adopted in generative evaluation.
Hu et al. [55]	2023	Text + Image	Proposes TIFA, a question-driven faithfulness evaluation pipeline decomposing prompts into structured QA verification tasks.	Provides interpretable, question-driven evaluation of text-to-image faithfulness beyond scalar similarity.
Goyal et al. [56]	2023	Text + Image	Introduces GENEVAL, an object-centric evaluation framework leveraging detection and attribute verification for compositional alignment assessment.	Provides fine-grained compositional alignment evaluation exposing relational reasoning failures.
Part II				
Reference	Year	Modalities/Data	Methodological Core	Key Contribution
<i>Human Preference and Reward Models</i>				
Wu et al. [57]	2023	Text + Image	Introduces HPS v2, a CLIP fine-tuned human preference scoring model trained on large-scale pairwise annotation data.	Establishes a large-scale human-aligned benchmark enabling more reliable preference-based model comparison.
Xu et al. [58]	2023	Text + Image	Introduces ImageReward, a learned reward model trained on expert preference comparisons for automatic scoring and diffusion optimization.	Enables automatic human-aligned scoring and direct diffusion model optimization via learned reward modeling.
Kirstain et al. [59]	2023	Text + Image	Introduces PickScore, a CLIP-based preference prediction model trained on real-user pairwise comparison data.	Provides large-scale human-aligned benchmark for ranking and evaluating T2I models.

Table 2. Cont.

Part II				
Reference	Year	Modalities/Data	Methodological Core	Key Contribution
<i>Datasets and Benchmarks</i>				
Li et al. [60]	2023	AGIQA-3K	Human-annotated multi-dimensional benchmark dataset with MOS supervision and structured evaluation protocol.	Introduces a multi-dimensional AIGIQA benchmark enabling systematic evaluation of perceptual and alignment quality.
Li et al. [61]	2024	AIGIQA-20K	Large-scale dataset expansion incorporating artifact-level annotations and structured training splits.	Expands large-scale supervision to improve robustness and generalization in AIGIQA modeling.
Yuan et al. [62]	2024	PKU-AIGIQA-4K	Unified perceptual dataset spanning text-to-image and image-to-image generative settings.	Extends perceptual benchmarking to multiple generative settings for broader QA evaluation.
Liu et al. [63]	2024	NTIRE 2024 AIGC Quality Assessment benchmarking framework with standardized tracks and evaluation criteria.	Establishes benchmarking standards and comparative baselines for AIGC quality assessment.	
<i>Surveys and Context (AIGIQA Landscape)</i>				
Wang et al. [64]	2025	Text + Image	Proposes a taxonomy-driven structured review categorizing text-to-image metrics into compositional and general quality dimensions.	Systematizes distribution-based, semantic, and preference-based metrics with actionable evaluation guidance.
Ghildyal et al. [65]	2024	Text + Image+Video	Presents an analytical synthesis framework reviewing architectures, datasets, and evaluation paradigms in generative quality prediction.	Synthesizes emerging research directions and challenges in generative quality assessment.
Tian et al. [66]	2025	Text + Image	Introduces a task-specific quality modeling framework tailored to visual communication contexts.	Frames quality assessment within visual communication applications, linking perception and authenticity.
Vinothkumar et al. [67]	2024	Text + Image	Provides a structured architectural analysis of diffusion-based text-to-image generation pipelines and conditioning mechanisms.	Provides foundational architectural context informing evaluation challenges in T2I systems.

4.2. Overview of Quality Assessment and Performance Metrics for AI-Generated Images

The evolution of quality assessment and performance metrics for AI-generated images reflects a clear shift from traditional distortion-based image quality assessment toward multi-dimensional, human-centred evaluation frameworks. Early objective measures such as Inception Score, Fréchet Inception Distance, and Kernel Inception Distance are widely used for evaluating generative model performance at the distribution level, but remain limited because they provide largely single-dimensional, group-level evaluation and are not ideal for instance-level quality assessment or for capturing semantic alignment with user intent. Large-scale MOS datasets underpin this shift. For example, AIGIQA-20K provides subjective scores for perceptual quality and text–image alignment, with the distinct contribution of annotating five specific artifact types such as blur, noise, compression, color distortion, and structural distortion, thus allowing a more diagnostic and fine-grained evaluation of AI-generated images.

A recent survey on quality metrics for text-to-image generation [64] synthesizes existing evaluation strategies into three primary categories: distribution-based realism metrics, cross-modal semantic alignment metrics, and human-centred perceptual/preference modeling approaches. The survey emphasizes that no single metric adequately captures perceptual realism, compositional reasoning, and user satisfaction simultaneously, reinforcing the need for hybrid multi-dimensional evaluation frameworks. This taxonomy aligns with the progression observed in contemporary AIGIQA research and further justifies the integration of the perceptual, semantic, and preference-aware evaluation paradigm.

Li et al. [60] introduced AGIQA-3K, a multi-dimensional benchmark specifically designed for evaluating AI-generated images. Their study demonstrated that distortion-centric IQA metrics are inadequate for AIGI, as high visual realism may still be accompanied by low authenticity or poor prompt faithfulness. By organizing quality assessment around perceptual visual quality, authenticity/naturalness, and text–image consistency, AGIQA-3K established a reliable foundation for training and validating AIGIQA models and to compare objective metrics against human judgments. Moreover, the benchmark highlights the growing importance of human-centred annotation protocols, including MOS and pairwise preference designs, as well as standardized correlation-based evaluation practices, to ensure reproducible and fair comparison across generative models and prompts.

In parallel, Tian et al. [66] introduced AIGCIQA2023, further reinforcing authenticity as a core quality dimension and highlighting that visually appealing generations may still be perceived as synthetic, motivating evaluation protocols and metrics that explicitly account for human judgments of naturalness alongside alignment and perceptual quality. Following these dataset-driven advances, Zhou et al. [45] proposed CIA-Net, a cross-modal interactive attention framework based on prompt-awareness across the modalities that jointly encodes the text prompt and the generated image using CLIP and outputs three complementary quality indicators: consistency, visual quality and authenticity, thereby supporting multi-dimensional evaluation aligned with subjective scoring practice.

In the same vein, Yuan et al. [47] introduced TIER, a text–image encoder-based regression approach that explicitly incorporates prompt information by extracting text features and image features and combining them through concatenation of features to predict quality, while also noting limitations in the deeper intermodal interaction. Beyond full-reference, prompt-aware approaches, other studies explore no-reference AIGIQA models that predict quality directly from the generated image without access to the prompt, leveraging deep feature statistics or generative-model priors to support real-world scenarios where reference information is unavailable.

At the benchmark level, Liu et al. [63] reported the NTIRE 2024 AI-Generated Content Challenge Quality Assessment, which standardized the evaluation using correlation-based criteria (SRCC/PLCC) computed against MOS with controlled datasets split between AIGIQA benchmarks, including both text-to-image and image-to-image datasets such as PKU-I2IQA. Together, these developments illustrate how contemporary AIGI evaluation integrates subjective quality assessment, objective model-based predictors, artifact-based analysis, and standardized benchmarking protocols to comprehensively assess both perceptual realism and semantic fidelity of AI-generated images.

4.3. Objective Quality Assessment

Objective metrics for AI-generated images (AIGIQA) are automatic and reproducible measures designed to quantify the quality of images produced by generative models by extracting measurable evidence from the generated image and, in many cases, its associated text prompt and mapping this evidence to a scalar quality score that enables consistent comparison between models, prompts, and datasets [60,62,63,65].

Objective evaluation is essential because AIGI may appear visually convincing while still containing non-natural artifacts, unrealistic structures, or semantic inconsistencies with the conditioning prompt that are not immediately apparent from visual inspection alone [52,60]. Consequently, recent AIGIQA datasets, benchmarks, and challenge reports increasingly treat image quality as an inherently multi-dimensional concept, requiring evaluation frameworks that jointly reflect perceptual fidelity, statistical realism, and prompt faithfulness [60,62,63,68]. In practice, the literature commonly organizes objective evaluation into three complementary categories.

4.3.1. Traditional Handcrafted No-Reference IQA Metrics

Traditional handcrafted no-reference image quality assessment (NR-IQA) metrics estimate visual quality by modeling deviations from natural scene statistics (NSS) and do not require access to a reference image. Among these approaches, BRISQUE is a representative metric that characterizes statistical regularities of locally normalized luminance coefficients to quantify generic photographic distortions, such as noise, blur, and compression artefacts [69]. Due to their low computational complexity and interpretability, handcrafted NR-IQA metrics have been widely adopted as baseline evaluators in image quality research.

However, recent studies on AI-generated image quality assessment demonstrate that these metrics exhibit limited effectiveness for AIGI, as they are primarily calibrated on distortions arising from natural image acquisition pipelines rather than artifacts specific to generative models. Experimental results reported on large-scale AIGIQA datasets, including AGIQA and AGIQA-20K, as well as findings from the NTIRE 2024 AI-Generated Content Challenge Quality Assessment, indicate a weak correlation between handcrafted NR-IQA scores and subjective human judgments when evaluating specific generative failures, such as structural inconsistencies, texture hallucinations, and unnatural object boundaries [60,63]. Consequently, handcrafted NR-IQA metrics are generally regarded as auxiliary baselines rather than reliable indicators of perceptual quality for modern AI-generated images.

4.3.2. Deep Feature-Based Metrics

To better capture perceptual characteristics beyond low-level distortions, deep feature-based metrics have been increasingly used for the objective evaluation of AI-generated images. These methods utilize pre-trained deep neural networks to extract high-level representations that encode semantic and perceptual information more closely aligned with human visual perception. In the literature, deep feature-based approaches can be broadly categorized into two types. Distribution-level metrics, such as the widely used Fréchet Inception Distance (FID) and Kernel Inception Distance (KID), evaluate the realism and diversity of generated image sets by comparing deep feature distributions between real and synthetic images, thereby reflecting global statistical properties of generative models.

In contrast, instance-level predictors, including representative methods such as GIQA and MANIQA-SC, estimate the perceptual quality of individual images by learning regression models over deep feature statistics, allowing for a fine-grained quality assessment at the image level [61,65]. Extensive evaluations of benchmark datasets such as AGIQA and AGIQA-20K, together with the results of the NTIRE 2024 challenge, show that deep feature-based metrics consistently outperform handcrafted NR-IQA baselines in terms of correlation with human subjective scores, particularly with respect to perceptual realism and authenticity [60,61]. Nevertheless, their performance can be influenced by factors such as feature extractor selection and domain mismatch, highlighting the need for complementary evaluation strategies.

4.3.3. Semantic Alignment Metrics

Although deep feature-based metrics effectively capture perceptual realism, they do not explicitly assess whether a generated image is semantically consistent with its conditioning input, typically a text prompt. To address this limitation, semantic alignment metrics have emerged as a crucial component of objective AIGI evaluation. Cross-modal similarity-based approaches, such as Clip Score, compute the similarity between image and text embeddings derived from vision-language models, providing a reference-free measure of prompt consistency [54]. Beyond similarity scoring, faithfulness-orientated methods, such as TIFA, evaluate semantic correctness by verifying objects, attributes,

and relationships implemented in prompt using structured question answering, allowing the detection of fine-grained compositional errors [55].

Recent advances extend similarity-based alignment metrics toward preference-driven reward modeling. Image Reward [58] introduces a human-aligned reward model trained from large-scale pairwise preference annotations, enabling instance-level scoring that correlates better with human judgement than embedding similarity metrics alone. Unlike Clip Score, which measures cosine similarity in a joint embedding space, Image Reward learns a ranking function optimized to predict which generated image humans prefer given a prompt, thereby capturing nuanced perceptual and compositional factors beyond semantic overlap.

Similarly, Pick Score, derived from the Pick-a-Pic dataset [59], trains a CLIP-based scoring function on more than 500,000 real-user preference comparisons. Experimental results demonstrate a stronger correlation with human rankings than traditional distribution-based metrics such as FID, highlighting the growing importance of preference-aware evaluation in text-to-image evaluation.

More recent prompt-aware AIGIQA models, including AMFF-Net and SF-IQA, integrate semantic alignment with perceptual quality prediction using cross-modal feature fusion and quality–similarity integration strategies, thus improving agreement with human judgments [45,49]. Subjective–objective comparative studies further confirm that incorporating semantic alignment substantially enhances the reliability of the evaluation, particularly to identify failures that purely visual metrics overlook [51,52].

4.3.4. Fine-Grained Compositional Evaluation

Although embedding-based similarity metrics evaluate global semantic correspondence, they often fail to detect object-level compositional errors. Goyal et al. [56] introduce an object-focused evaluation framework that decomposes requests into explicit entity, attribute, and relational components and assesses whether each is properly rendered in the generated image. By systematically varying compositional complexity, GENEVAL reveals that many state-of-the-art generative models struggle with spatial relations and multi-object reasoning despite achieving strong FID or Clip Score results.

This object-centric evaluation protocol highlights a critical limitation of conventional metrics: high perceptual realism does not guarantee compositional correctness. Consequently, fine-grained evaluation frameworks are increasingly necessary for diagnosing reasoning failures in generative systems.

4.3.5. Preference-Based and Reward Modeling Metrics

Preference-based evaluation has emerged as a powerful alternative to distribution-level metrics by directly modeling human comparative judgments. Human Preference Score v2 (HPS v2) extends this paradigm by constructing a standardized benchmarking protocol for text-to-image models using curated human ranking data on diverse prompts [57]. Unlike scalar MOS aggregation, pairwise preference modeling enables finer discrimination between visually similar generations and is particularly effective for evaluating stylistic and compositional fidelity.

These approaches shift the evaluation paradigm from similarity estimation toward learnt reward modeling, aligning quality prediction with human satisfaction rather than statistical realism alone. As demonstrated in large-scale studies, reward-based metrics consistently outperform FID and embedding-based alignment measures in predicting user preference trends across generative models.

4.4. Subjective Image Quality Assessment and Human Studies

Li et al. [60,61] consistently position subjective evaluation as the ground truth for AI-generated image (AIGI) quality, because it directly measures what human observers actually perceive rather than what a model estimates numerically. In the AIGI setting, quality degradations are not limited to classical distortions such as blur or noise but also include generative-specific failures such as unnatural textures, implausible object structures, compositional artefacts, and prompt-related semantic inconsistencies, including missing objects, incorrect attributes, and violated relationships.

Liu et al. [63] further emphasize that subjective scores typically in the form of Mean Opinion Scores (MOS) or pairwise preference labels are used in three primary ways in the literature: (i) to construct large-scale AIGIQA datasets for supervised learning and benchmarking; (ii) to benchmark how well objective quality metrics correlate with human perception; and (iii) to train learning-based quality predictors and ranking models that aim to approximate human judgement on scale. Consequently, this subsection synthesizes how perceptual quality in AI-generated images is defined in terms of human visual perception dimensions, how subjective AIGIQA studies formalize these dimensions, how MOS is collected through standardized protocols, how pairwise comparisons support preference learning, and which reliability and bias-control mechanisms are reported to ensure trustworthy annotations.

Beyond MOS aggregation, large-scale user-driven preference datasets have recently expanded the scope of subjective supervision. The Pick-a-Pic dataset contains more than 500,000 prompt–image preference pairs collected from real users via an interactive generation interface, enabling large-scale reward modeling. In contrast to a controlled laboratory-style MOS collection, this approach captures authentic user intent and stylistic variation [59].

Similarly, Image Reward uses extensive pairwise annotations to train a reward model optimized for ranking consistency, demonstrating that preference learning provides stronger predictive power for model evaluation than absolute scoring schemes. These findings suggest that preference-based supervision may complement or, in some cases, replace traditional MOS-based evaluation in large-scale generative assessment.

4.4.1. Perceptual Quality

Perceptual quality in AI-generated images (AIGI) describes the degree to which generated content aligns with human judgments of visual realism, naturalness, aesthetic appeal, and structural coherence, capturing holistic viewing experience rather than isolated pixel-level fidelity. Unlike natural images, whose perceptual degradation is primarily associated with acquisition or compression artefacts, AIGI exhibit distinct failure modes such as unnatural textures, spatial incoherence, and semantic inconsistencies, rendering traditional distortion-orientated IQA measures insufficient. Consequently, recent research consistently positions human subjective evaluation as the ground truth for perceptual quality assessment, typically operationalized through controlled experiments that collect Mean Opinion Scores (MOS) reflecting aggregated human perception.

Li et al. [60,61] further emphasize that MOS obtained through controlled subjective experiments constitute the ground truth for both benchmarking objective quality metrics and training-based perceptual quality predictors.

Large-scale benchmarks such as AIGIQA-3K and AIGIQA-20K explicitly incorporate perceptual quality annotations and demonstrate that classical no-reference metrics based on natural scene statistics (e.g., BRISQUE) exhibit a weak correlation with human judgments when applied to generative imagery [60]. In particular, the AIGIQA-20K dataset was constructed by collecting extensive MOS labels across images generated by diverse text-to-

image models and sampling strategies, showing that perceptual quality is influenced not only by low-level visual realism but also by high-level semantic coherence and the internal plausibility of the generated scene [61].

Yu et al. [49] further demonstrate through large-scale subjective evaluation that human perception is inherently multi-dimensional, as participants rate AI-generated images jointly with respect to visual quality and text–image consistency.

To improve reliability and reduce observer bias, the field has adopted refined subjective evaluation protocols. These include rater training, randomized image presentation, standardized display resolution, and constrained session duration to mitigate visual fatigue, together with statistical post-processing procedures such as z-score normalization and subject outlier rejection according to the ITU-R BT.500 recommendations [63]. In addition to absolute rating schemes, pairwise comparison tests are often more sensitive than absolute ratings, and preference learning schemes are especially effective when evaluating subtle or stylistic differences [49]. These human-derived perceptual scores are subsequently used to benchmark learning-based perceptual quality predictors, where rank and linear correlation measures (SRCC and PLCC) quantify alignment with human opinion [63]. Overall, the surveyed literature converges on the view that perceptual quality in AIGI is inherently multi-dimensional and human-centric, requiring dedicated subjective measurement protocols and learnt perceptual models that explicitly account for the unique characteristics of AI-generated visual content rather than relying on assumptions inherited from natural-image quality assessment.

4.4.2. Human Visual Perception Dimensions

Zhou et al. and Liu et al. [45,63] demonstrate that human perception of AI-generated images is inherently multi-dimensional and cannot be adequately captured by a single scalar notion of image quality. As a result, subjective studies typically decompose perception into several complementary dimensions. Perceptual or visual quality reflects low- and mid-level attributes such as sharpness, visible artefacts, color fidelity, contrast, structural coherence, naturalness, and overall viewing comfort. Authenticity or realism evaluates whether an image appears photorealistic and indistinguishable from real-world imagery rather than recognisably synthetic [51].

Several studies also incorporate aesthetic quality or artfulness, particularly for artistic or stylized generation tasks, where rating agencies evaluate visual appeal, stylistic consistency, and creative expressiveness [68]. Semantic or prompt faithfulness, emphasized by [49,55], measures whether objects, attributes, spatial relations, and scene composition correctly reflect the input prompt. Importantly, user studies in Hu et al. and Chen et al. [51,55] implicitly validate the separability of these dimensions by instructing annotators to score them independently and reporting divergent trends between dimensions, indicating that visually appealing images may still be semantically incorrect or implausible. These dimensions therefore define the subjective labels collected in AIGIQA datasets and represent the perceptual targets that objective metrics and learning-based predictors aim to approximate.

4.4.3. MOS and Subjective Protocols

Chen et al. [51] report that the Mean Opinion Score (MOS) remains the most widely adopted mechanism for aggregating subjective judgments of AI-generated image quality and serves as the primary perceptual ground truth in existing benchmarks. In typical MOS-based protocols, participants rate images using discrete or quasi-continuous scales (e.g., 1–5 or 0–5), sometimes allowing half-step increments or continuous sliders to capture finer perceptual differences. Depending on the design of the dataset, MOS can represent a

single overall quality score or multiple aspect-specific scores, such as separate MOS values for perceptual quality, aesthetics, and text–image alignment [45].

Large-scale benchmark datasets further illustrate protocol choices; see, for example, Liu et al. [63] report that AIGIQA-20K contains 20,000 AI-generated images from 15 text-to-image models, with MOS labels collected from 21 subjects. Viewing conditions generally follow established subjective IQA practices, including image resizing to fixed resolutions, controlled display assumptions, and randomized presentation order. Yuan et al. [62] explicitly reference the ITU-R BT.500 guidelines, demonstrating how classical subjective IQA standards are adapted to generative imagery. MOS values are computed as the mean between raters, with some studies reporting variance or applying rater filtering to improve robustness, producing stable perceptual targets for benchmarking objective metrics and training supervised quality predictors.

4.4.4. Pairwise Comparison and Preference Learning

Hu et al. [55] note that pairwise comparison is widely adopted in image generation evaluation because relative judgments are cognitively easier and often more consistent for human observers than absolute scores. In pairwise protocols, participants are presented with two images typically generated from the same prompt but produced by different models or sampling strategies and asked to select the image that best satisfies a criterion such as realism, semantic alignment, or aesthetic appeal. Image pairs are commonly presented side by side in randomized order to reduce position and ordering bias.

Chen et al. [68] provide a concrete example of how such annotations support preference learning by adopting a learn-to-rank formulation to model relative artefacts rather than absolute scores. Pairwise preference data are also used to evaluate automatic metrics by measuring the agreement with human choices. Compared to MOS, which provides an absolute quality estimate, pairwise comparison captures relative perceptual differences; consequently, many AIGIQA benchmarks treat MOS and pairwise preferences as complementary signals to model human judgement [49].

4.4.5. Reliability and Bias Control

Hu et al. [55] provide concrete evidence of how reliability is quantified and improved in subjective AIGI evaluation by reporting between-annotator agreement using Krippendorff's alpha: $\alpha = 0.67$ for Likert-based holistic faithfulness judgments and $\alpha = 0.88$ for structured VQA binary annotations (both values reported in [55]). The substantial gap between these two figures demonstrates that decomposing faithfulness evaluation into structured binary questions about specific prompt attributes substantially stabilizes inter-annotator agreement relative to holistic Likert scoring—a design principle with direct implications for benchmark construction. The study also employs adjudication via majority voting when a third annotator resolves binary disagreements. Beyond agreement metrics, subjective AIGI evaluation is susceptible to rater fatigue, inconsistent calibration, order effects, prompt-induced context bias, and display or device variability.

To mitigate these risks, reviewed studies commonly randomize image presentation order, particularly within prompt groups, and split evaluations into shorter sessions with enforced breaks Zhou et al., Liu et al. [45,63] further demonstrate that subjective scores vary systematically with prompt duration and generation settings, highlighting AIGI-specific biases that require careful protocol design and balanced sampling. Attention cheques, rating qualification procedures, and detailed scoring guidelines such as those used by Hu et al. [55] to clarify how missing or incorrect elements should be judged are used to improve annotation consistency. Despite these controls, subjective evaluation remains expensive and difficult to scale, motivating hybrid pipelines in which human annotations

are strategically used to validate and guide objective or learning-based metrics rather than replace them entirely [52].

4.5. Learning-Based Perceptual Quality Prediction

Learning-based perceptual quality predictors aim to map AI-generated images (AIGI) to human-perceived quality scores, e.g., mean opinion scores (MOS), by training regressors on large annotated AIGIQA datasets [62]. A key distinction in recent AIGIQA research is whether the predictor relies on image-only signals, such as visual artefacts, realism, and aesthetics, or is prompt-aware, explicitly modeling text–image alignment and semantic faithfulness. This distinction is critical because many failure modes in AI-generated images, such as missing objects, incorrect attributes, or weak compositional binding, are semantic in nature, rather than purely low-level distortions. Li et al. [60,61] further confirm that prompt-aware modeling is essential for accurately reflecting human perceptual judgments, as many dominant AIGI degradations arise from semantic inconsistency rather than visual artefacts alone.

4.5.1. Image-Only Predictors

Early learning-based perceptual quality predictors predominantly adopt a no-reference, image-only setting, in which perceptual attributes are inferred solely from the generated image. These methods typically leverage deep visual encoders, such as convolutional neural networks or Vision Transformers, followed by regression heads trained to predict subjective quality scores. Image-only predictors focus primarily on modeling visual realism, texture fidelity, structural coherence, and aesthetic quality.

Representative approaches include image-centric regression models that encode deep visual representations and directly learn mappings to perceptual quality scores [47]. Experimental results consistently demonstrate that such learning-based image-only predictors substantially outperform traditional no-reference IQA methods based on natural scene statistics when applied to AI-generated imagery [60,61]. However, because these methods ignore the generative intent expressed in the prompt, they are inherently limited in assessing semantic correctness and prompt compliance.

4.5.2. Prompt-Aware Models

To overcome the limitations of image-only predictors, recent studies increasingly adopt prompt-aware perceptual quality prediction, in which both the generated image and its corresponding text prompt are jointly modeled. These approaches typically rely on pre-trained vision language encoders to embed images and prompts into a shared semantic space, followed by specialized modules that evaluate visual quality together with text–image alignment. Zhou et al. [45] introduce CIA-Net, which incorporates a cross-modal interactive attention mechanism to explicitly model intermodal dependencies between visual features and text embeddings. Leaving for fine-grained interaction between image and prompt representations, CIA-Net jointly predicts perceptual quality, text–image consistency, and authenticity.

Zhang et al. [48] propose AMFF-Net, which decomposes perceptual evaluation into multiple dimensions, including visual quality, authenticity, and consistency. Their method employs a multi-scale feature extraction strategy combined with adaptive feature fusion to capture perceptual cues across different spatial resolutions while incorporating prompt-conditioned semantic alignment. Yu et al. [49] present SF-IQA, which integrates a visual quality perception branch with a semantic similarity assessment branch and fuses both components into a unified perceptual quality score, reflecting the inherently multi-dimensional nature of human judgement.

Experimental evaluations on widely used benchmarks, including AGIQA-3K and AIGIQA-20K, show that prompt-based learning-based predictors consistently outperform traditional no-reference IQA methods based on natural scene statistics and image-only learning models [49,60,61]. Collectively, the reviewed literature reveals a clear shift toward data-driven, learning-based perceptual quality predictors trained on large-scale human subjective annotations while also highlighting their dependence on the quality, diversity, and reliability of the subjective datasets and benchmarking protocols used for training and evaluation.

4.6. Benchmarking Protocols

Benchmarking protocols play a critical role in evaluating the effectiveness and reliability of perceptual quality assessment methods for AI-generated images, providing standardized procedures to compare objective metrics and learning-based predictors against subjective human judgments. Li et al. [60,61] describe the use of AGIQA-3K and AIGIQA-20K as benchmark datasets with subjectively annotated quality scores for systematic evaluation.

Benchmark experiments are commonly conducted using predefined train–test splits with safeguards against content leakage, ensuring that images generated from the same prompts do not appear in both training and testing sets. Zhang et al. [48] report performance primarily using correlation-based criteria, including SRCC and PLCC, to quantify the consistency of the ranking and the agreement with the mean opinion scores. In large-scale benchmarking efforts such as the NTIRE 2024 AIGC Quality Assessment Challenge, Liu et al. [63] further divide the datasets into training, validation, and testing subsets and rank the models using combined SRCC–PLCC scores after nonlinear regression mapping. Several studies also highlight the importance of cross-dataset evaluation to assess generalization performance and reveal dataset-specific biases, as discussed by [61].

Limitations of Existing Measures

Despite rapid progress, existing AI-generated image evaluation measures (AIGI) exhibit persistent limitations that constrain their reliability, interpretability, and real-world applicability. Many widely used objective metrics were originally developed for natural-image distortions or distribution-level generative evaluation and therefore do not capture generative-specific failure modes in modern text-to-image systems, including semantic inconsistencies, spatial incoherence, implausible object structures, and prompt-related attribute binding [60,61]. Consequently, conventional no-reference IQA baselines based on natural scene statistics, while computationally efficient and interpretable, are generally treated as auxiliary indicators due to their weak correlation with human judgments in generative settings [63].

Deep feature-based objective measures present a second limitation. Distribution-level metrics such as FID and KID provide useful model-level comparisons but remain insensitive to instance-level perceptual defects and do not directly quantify prompt faithfulness or semantic correctness [60,61]. Instance-level regressors improve correlation with MOS relative to handcrafted NR-IQA metrics; however, their performance depends on feature extractor choice and domain alignment, and they may still underperform on fine-grained compositional errors that are visually subtle yet semantically critical. Moreover, learning-based predictors are susceptible to dataset-specific priors, limiting generalization to new generators, unseen prompts, or novel rendering styles.

Semantic alignment and vision–language similarity metrics address part of this gap but introduce additional shortcomings. Similarity-based approaches capture coarse prompt–image alignment, yet struggle with fine-grained attribute binding, object counting, spatial relations, and compositional reasoning failures [55,60]. Faithfulness-orientated verification

improves interpretability but depends on the design of the question and can overlook perceptual artefacts that degrade realism without violating explicit prompt constraints [49,55]. Prompt-aware perceptual predictors improve agreement with human judgments yet remain constrained by the representational limits of underlying vision–language encoders and imperfect multi-modal reasoning [45,48,49].

Fine-grained compositional evaluation further exposes structural weaknesses in current measures. Goyal et al. [56] demonstrate that generative models can achieve strong global similarity or perceptual scores while failing object-level reasoning tasks such as attribute binding and spatial relations. These results show that perceptual realism or embedding similarity does not guarantee compositional correctness. Although GENEVAL provides an interpretable diagnostic protocol, it functions primarily as a benchmark rather than as a unified perceptual quality measure.

Preference-based reward modeling introduces additional considerations. Methods such as Image Reward and HPS v2 show a stronger correlation with human satisfaction than similarity-based metrics [57,58]. However, reward models are highly dependent on the diversity and representativeness of preference data and can encode stylistic or dataset-specific biases. Although effective for ranking, they typically produce scalar outputs without disentangling perceptual realism, semantic correctness, and aesthetic preference, limiting interpretability.

Taken together, current AIGI evaluation paradigms remain dimension-specific and fragmented, motivating the development of integrated multi-dimensional frameworks that jointly model perceptual fidelity, compositional reasoning, semantic faithfulness, and human preference within standardized benchmarking settings.

A further and underexamined layer of concern involves the *validity of the benchmarking infrastructure* itself, independent of the metrics it hosts. Five inter-related threats have received insufficient systematic attention in the retained literature. First, *CLIP encoder bias*: CLIP-derived similarity metrics and CLIP-fine-tuned reward models inherit the biases of CLIP’s training distribution, which over-represents Western photographic aesthetics and English-language semantic categories; this means that generation models scoring highly on CLIP-derived metrics may nonetheless fail to represent diverse visual cultures or to satisfy prompts whose semantic content is underrepresented in CLIP’s vision–language pretraining [65]. Second, *prompt leakage*: if the prompts used to construct a benchmark overlap with the training distribution of the generative models being evaluated, benchmark scores will reflect memorization or near-memorization rather than generalization; this is an especially acute concern for models that cannot disclose the full scope of their pretraining data. Third, *reward hacking*: when generative models are fine-tuned against human preference reward models [57,58], the resulting optimization pressure can move the model distribution toward reward-model-preferred imagery that is not genuinely preferred by independent human raters—a form of distributional overfitting to the reward proxy. Fourth, *benchmark saturation*: as models advance, several existing benchmarks exhibit score compression near their upper bounds; on GENEVAL [56] and related compositional benchmarks, recent frontier models already approach ceiling performance on simpler attribute-binding tasks, reducing their discriminative value for ranking state-of-the-art systems and motivating the development of next-generation, higher-complexity diagnostic suites. Fifth, *annotation instability*: human preference is temporally non-stationary—aesthetic expectations evolve with exposure to increasingly high-quality generated content—and cross-annotator agreement on subtle quality dimensions such as naturalness and aesthetic appeal is measurably lower than agreement on binary correctness judgments [55]; datasets collected at different time points or with different annotator pools may therefore not be directly comparable. Addressing these benchmark validity threats requires community-level coordination on

auditable prompt curation, periodic calibration studies, and the development of evaluators that are explicitly robust to the biases of their underlying vision–language encoders.

4.7. Section-Specific Limitations

The quality assessment corpus of 23 studies is concentrated in a small set of specialized benchmarks, most notably AGIQA-3K, AIGIQA-20K, and the NTIRE 2024 AIGC Quality Assessment Challenge, which shapes the extent to which cross-study comparisons can be generalized. Annotation protocols, prompt diversity, and the mix of generative models used to produce stimuli differ substantially across these benchmarks, which limits direct comparability of correlation coefficients and ranking outcomes reported against different ground-truth sets. Several retained studies also combine benchmark proposal with internal evaluation, meaning that the same data collection conditions underpin both training and test conclusions. Taken together, these characteristics mean that the synthesis in this section reflects the dominant methodological directions and recurrent shortcomings of the retained evidence, rather than a calibrated ranking of methods across a unified evaluation standard.

5. Applications of AI-Driven Image Generation

In this section, we survey research articles that focus on the application of generative-image AI to various domains. The technology has begun to reshape professional practice across a wide range of applications from general product development and architectural design to fashion, graphic arts, storytelling, and 3D synthesis. By integrating specialized frameworks such as multi-agent systems, theory-guided conceptual models, and modular 2D-to-3D pipelines, generative AI is expanding the boundaries of technical precision, realism, and expressive capability. Yet, the research literature also highlights significant challenges regarding emotional dynamics, alignment, and the need for discipline-specific expertise. Together, these works illustrate the growing maturity of AI-driven image generation as a transformative, yet complex, partner.

5.1. Section-Specific Selection and Thematic Grouping

Within the global review workflow described in Section 2, the applications branch contained 308 records after thematic assignment. Section-specific screening then excluded papers not centred on a concrete application domain, as well as works in which image generation was incidental or only marginally analyzed, reducing the pool to 103 candidates. A final application-oriented refinement retained studies that addressed reasonably broad domains and reported substantive experimental or evaluative evidence in domain-specific scenarios. This yielded the final corpus of 29 articles analyzed in this section.

The 29 retained articles were clustered into the following thematic categories based on the specific application domains they addressed: (1) *General Design*, (2) *Architectural Design*, (3) *Fashion Design*, (4) *Graphic Art*, (5) *Storytelling*, (6) *3D Scene Generation*, and (7) *3D Character Generation*. Each article was analyzed in terms of its methodological approach, the specific AI-driven image generation techniques employed, the methods used for evaluation, and the reported outcomes. Table 3 presents the retained study characteristics, and the narrative synthesis then compares the reported application evidence across these domain clusters.

5.2. General Design

Across contemporary design research, generative-image AI is emerging as a powerful tool to facilitate creativity, communication, and co-creation in the formation of ideas and concepts. The research articles in this category illustrate the breadth of this transformation: from efforts to enhance designer-user co-creation to tools that streamline design exploration to tools that use AI-assisted image generation to enhance transparency and clarity in design feedback.

Table 3. Summary of representative applications of AI-driven image generation Part I and Part II.

Part I				
Reference	Year	Modalities/Data	Methodological Core	Key Contribution
General Design				
He et al. [70]	2024	Text→image (Midjourney); co-creative workshop with 6 designer–user dyads; AI-generated visuals	Iterative co-creation model integrating generative-image AI across five phases to surface needs	Shows AI imagery enriches dialogue and reveals tacit needs but can bias perceptions, requiring designer mediation
Choi et al. [71]	2024	Reference images; keyword suggestions; low-fidelity sketch generations; user study (n = 16)	System extracts attributes from references, suggests keywords, and generates conceptual recombinations as sketches	Increases idea volume and self-reported creativity versus a baseline system
Chen et al. [72]	2024	Generative models + design theories (5W1H, FBS, Kansei); interactive mind-map interface; user studies (n = 30, 10)	Theory-guided multi-stage generation that makes reasoning steps explicit for transparency and control	Outperforms baselines with more rigorous, emotionally resonant concepts and improved designer agency
Duan et al. [73]	2026	Generative artificial intelligence; feature composition from reference products; two-phase user study (n = 8, 24)	Structured workflow that lets novices extract and compose features from referenced products	Yields more diverse concepts and a more immersive, engaging design experience than baselines
Chen et al. [74]	2024	Text editor + CLIP-based viewpoint suggestion; text+scribble, Grab’n Go, inpainting; two user studies (n = 14, 8)	Combines automatic viewpoint selection with lightweight editing tools to generate intent-focused reference images	Beats sketches/web search for communicating design intent; ~66.7% images rated effective
Architectural Design				
Lu et al. [75]	2025	Text→image; cultural knowledge base; multi-agent system; single case study	Knowledge-driven modular MAS with closed-loop validation coordinated by a scheduler	Improves quality, cultural fidelity, and intent alignment while lowering barriers for designers
Guida [76]	2023	Text→image→three-dimensional workflows; CLIP-guided diffusion (Stable Diffusion, DALL-E 2); site/environment/ building-code constraints	Three workflows including conditioning generation on reconstructed 3D massing models to embed constraints	Shows flexible integration of AI that balances creative exploration with contextual precision and fosters trust
Paananen et al. [77]	2023	Text-to-image; lab study with students; questionnaires, prompt analysis, group interviews	Empirical assessment of T2I tools’ role in early-stage ideation and workflows	T2I can spark discovery when paired with constraints; highlights need for better training and features
Fashion Design				
Guo et al. [78]	2025	DreamBooth-tuned Stable Diffusion; LRM 3D reconstruction; CATIA flattening; CLO 3D virtual fitting	2D→3D→2D pipeline coupling diffusion generation with triplane-based reconstruction and pattern flattening	Produces accurate 2D sewing patterns with deviations < 0.5 cm in virtual fitting
Rizzi et al. [79]	2025	Midjourney, ChatGPT (version not reported), Runway in studio lab; reinterpretation of historical jacket; student teams	Qualitative study mapping modes of human-AI collaboration and layers of design knowledge	AI augments ideation/visual exploration but misses structural/material nuances; designer expertise remains vital
Sun et al. [80]	2024	Fashion product images; two-stage neural style transfer; custom 5-factor evaluation	MRF-based low-res optimization + full-res Gram loss with fashion-specific constraints (symmetry/gradients)	Delivers efficient, high-quality style transfer while preserving garment background
Graphic Art				
Ahn et al. [81]	2024	Stable Diffusion; reference-guided stylization; textual inversion	Specialized denoising with structural conditioning + context-aware prompt augmentation and dual guidance	Enables faithful style inversion with content preservation and fine-grained control
Azuaje et al. [82]	2023	Lyrics (DM-GAN) + audio emotion analysis; style bank + style transfer; user study (n = 35)	Dual-model pipeline blending lyric-driven synthesis with mood-guided style selection and transfer	>60% participants positive; accessible creation of emotionally aligned album art
Ko et al. [83]	2023	Systematic review (72 papers) + 28 artist interviews	Mixed-methods analysis deriving design roles and interface guidelines for LTGMs	Identifies 3 roles (automation, exploration, mediation) and proposes guidelines for controllable, customized tools
Almeda et al. [84]	2024	Spreadsheet interface with LLM functions; TTI prompt grids; lab (n = 12) + 2-week expert deployment (n = 5)	Spreadsheet-native composable functions to systematically vary prompts and map design spaces	Supports deliberate, interpretable exploration and sense-making in generative art workflows

Table 3. Cont.

Part I				
Reference	Year	Modalities/Data	Methodological Core	Key Contribution
Hod et al. [85]	2025	PSD/PSB assets; LLaVA prompts; LoRA on in-game assets; GPT-4o evaluator; pro designer study (n = 9)	Divide-and-conquer pipeline (layer extraction → bg generation → layout refinement → harmonization)	Achieves 50–85% efficiency gains while preserving visual quality
Part II				
Reference	Year	Modalities/Data	Methodological Core	Key Contribution
<i>Storytelling</i>				
Leininger et al. [86]	2025	Depth Mesh, Panoramic Meshes, Gaussian Splatting; Unreal Engine prototype; 15 industry experts	Prototype (EnVisualAIzer) integrating diffusion and Gaussian Splatting to test environment types	Empirically validates use cases and pinpoints usability/fidelity gaps
Fernandes et al. [87]	2024	Story keywords; Stable Diffusion; user study (n = 7)	Keyword extraction pipeline to translate stories into image prompts for SDXL	Boosts creativity and engagement but needs clearer onboarding and more precise keywords
Shahriyar et al. [88]	2026	GPT-4o narratives + character sheets; MMDiT image gen; online platform; CLIP-based metrics	Three-module framework (Generation/Alignment/Image) with running state to enforce text-image consistency	Delivers consistent, personalized stories with high user satisfaction and fewer regenerations
Liang et al. [89]	2025	GPT-4 + Stable Diffusion; 12 UX students; dual workflows observed	Dual-model pipeline auto-segments narratives and renders stylistically consistent images with editable prompts	Reduces time and workload while improving storytelling quality and inspiration
Yi Chan et al. [90]	2024	Stable Diffusion 1.5 + ControlNet; CLIP encoder; GPT-3.5-based workflow; user study (n = 50)	Sketch-guided generation with Canny edges and LLM-driven narrative structuring	Accelerates character storyboard production and improves structural interpretation despite occasional mismatches
<i>3D Scene and Object Synthesis</i>				
Numan et al. [91]	2024	Images→3D submeshes; depth estimation; VLM/LLM prompts; diffusion inpainting; VR user study (n = 20)	Two-stage blend of reconstructed submeshes with language-guided inpainting to form navigable VR spaces	Improves comfort and task performance vs. baselines; needs better geometric fidelity
Epstein et al. [92]	2024	Text-to-three-dimensional scene generation; multiple NeRFs; score distillation; unsupervised	Optimizes multiple NeRFs across layouts so each learns a coherent object, enabling disentanglement	Enables editable, object-level control without extra supervision or annotations
Zhang et al. [93]	2024	Text prompt; diffusion init; DIBR multi-view support; progressive inpainting; depth-aware losses	Progressive view-by-view expansion combines diffusion inpainting with NeRF optimization	Generates photo-realistic, geometrically consistent 3D scenes from a single prompt
Pu et al. [94]	2024	Single panorama; SD with LoRA per scene; panoramic depth inpainting; 3D Gaussian Splatting	Back-project panorama to mesh, iteratively refine geometry/texture, then convert to 3DGS for dense views	Sets new benchmark for high-fidelity indoor reconstruction and novel view synthesis from one panorama
<i>3D Character Generation</i>				
Peng et al. [95]	2024	Single 2D image; Anime3D dataset; dual-UNet + transformer; NvDiffRast; user study (n = 21)	Pose canonicalization to A-pose with multi-view synthesis, followed by coarse-to-fine reconstruction and texture back-projection	Outperforms prior methods in geometry, texture quality, and multi-view consistency
Dong et al. [96]	2024	Text prompt; Skeleton-guided diffusion; 3D/NeRF	Progressive layer-wise generation that disentangles body and garments via stratified compositional rendering	Achieves superior structural disentanglement and realism, enabling virtual try-on and clothing transfer
Yamada et al. [97]	2025	Text prompt; Skeleton-guided diffusion; Instant Neural Graphics Primitives (Instant-NGP); ControlNet; anime-specific LoRA	Occlusion-aware skeleton extraction with anime-style normalization/inpainting loss within score distillation	Improves semantic alignment and stylistic fidelity for cel-shaded 3D avatars
Li et al. [98]	2026	Single image; SC-LRM with multiple semantic decoders; component-wise reconstruction	Semantic decomposition into body, clothing, hair, shoes with A-pose multi-view generation	Delivers cleaner geometry and modular assets suitable for rigging and reuse

5.2.1. Summary of Articles on General Design

In [70], He et al. examine an iterative model for integrating generative-image AI into early product development. Evaluated in a co-creative Midjourney workshop, the model showed that AI-generated visuals can strengthen user engagement and surface latent needs

but can also bias perceptions through their emotional impact. The study therefore frames AI imagery as both a support for co-creation and a source of interpretive risk.

In [71], Choi et al. present *CreativeConnect*, a system for extracting attributes from reference images, suggesting keywords, and generating low-fidelity sketch variations. Its main purpose is to support open-ended recombination rather than polished output. In a user study with 16 participants, it increased both idea volume and self-reported creativity relative to a baseline.

In [72], Chen et al. introduce *DesignFusion*, a framework that combines generative AI with classical design theories and Kansei Engineering to make reasoning steps more explicit. Across two user studies, the system outperformed baseline generative workflows in perceived rigor, emotional resonance, and clarity of interaction. The broader contribution is to show that theory-guided structure can improve both transparency and designer agency in AI-assisted design.

In [73], Duan et al. introduce *DesignFromX*, a support tool aimed at novice and consumer-led design. It structures the composition of features from reference products into a more accessible workflow. In a two-phase user study, participants produced more diverse concepts, incorporated more reference features, and reported a more engaging design experience than with baseline methods.

In [74], Chen et al. present *MemoVis*, a generative-AI-assisted editor for asynchronous 3D design critique. By combining viewpoint suggestion with lightweight visual editing tools, the system helps reviewers turn textual feedback into targeted reference imagery without requiring advanced modeling skills. Two user studies showed that MemoVis communicated design intent more effectively than sketches or web image search.

5.2.2. Controllability Requirements of General Design

Across all five papers, controllability is considered essential for transforming generative AI from an unpredictable inspiration tool into a practical co-design partner. Rather than functioning as autonomous generation systems, these tools operate across a spectrum of control tailored to specific task complexities and user expertise.

For early-stage ideation, He et al. [70] treat control as conversational, encouraging continuous dialogue and reflection during collaborative workshops to surface tacit user knowledge. Moving toward graphic ideation, CreativeConnect [71] structures control around four decomposed elements—subject, action, theme, and composition—and intentionally outputs low-fidelity sketches to prevent creative fixation while preserving designer authorship. For advanced engineering design, DesignFusion [72] rejects opaque, end-to-end generation by visualizing reasoning chains, allowing designers to inspect, modify, or regenerate every intermediate step. Finally, DesignFromX [73] and MemoVis [74] extend fine-grained steering to non-experts: the former replaces prompt engineering with direct, brush-based feature selection from reference images; the latter incorporates viewpoint-aware prompting, scribbles, and region-preserving modifiers so users without advanced 3D skills can communicate spatial critiques.

Collectively, the research establishes that effective design AI must offer iterative, interpretable, and context-sensitive exploration while maintaining human ownership over creative decisions.

5.2.3. Domain Constraints in General Design

The five papers collectively identify substantial constraints stemming from model opacity, dataset bias, limited domain reasoning, and the difficulty of representing tacit human knowledge. Several of the studies note that foundation models trained on generic

internet data often produce visually plausible but functionally invalid or semantically shallow outputs.

Chen et al. [72] point out that conceptual engineering tasks require understanding of feasibility, emotion, and manufacturing logic beyond what current LLMs and text-to-image systems can reliably provide, whereas Duan et al. [73] emphasize that novice consumers lack the specialized terminology needed to communicate detailed product preferences through text prompts alone. Additional constraints arise from spatial complexity, aesthetic bias, and usability barriers. Chen et al. [74] point to the challenge of aligning generated imagery with precise 3D viewpoints and contextual feedback requirements, while He et al. [70] warn that highly polished AI outputs may unintentionally narrow creative diversity through aesthetic fixation.

Across the papers, recurring issues include hallucinations, inconsistent prompt interpretation, limited support for engineering reasoning, generation latency, dependence on proprietary APIs, copyright ambiguity, and insufficient adaptation to specialized disciplinary needs. Together, the studies portray current generative AI systems as powerful but still constrained by incomplete contextual understanding and weak domain-specific reasoning capabilities.

5.2.4. Human-in-the-Loop Considerations for General Design

All five papers strongly frame generative AI as a collaborative assistant rather than a replacement for designers. Human users remain responsible for interpretation, critique, contextual reasoning, and final decision-making, whereas AI primarily supports ideation, visualization, and exploration. The systems are specifically engineered to facilitate human–AI and human–human collaboration rather than serving as mere prompt-operator interfaces.

He et al. [70] reframe AI as a mediator to support dialogue in participatory design workshops, helping users externalize tacit needs and reflect on possibilities. Choi et al. [71] demonstrate that deliberately presenting unfinished sketch-like outputs increases idea fluency because it actively invites human reinterpretation. Similarly, Chen et al. [72] embed human review directly into intermediate workflows by exposing design logic, allowing users to selectively filter requirements, select behaviors, and adjust structures, which significantly improves designer confidence and system transparency. To bridge the gap for non-designers, Duan et al. [73] show that exposing feature-level choices lowers user frustration, while Chen et al. [74] demonstrate how guided editing tools turn complex 3D reviews into actionable, co-creative tasks for novices.

Ultimately, across these studies, creative value emerges from an iterative loop where AI expands the search space and accelerates visualization, while humans retain sole responsibility for evaluating feasibility, addressing stakeholder needs, ensuring social relevance, and making final design decisions.

5.3. Architectural Design

Generative-image AI has also been applied to architectural design, offering new pathways for cultural fidelity, contextual responsiveness, and creative exploration. The research articles in this category include work on culturally informed visualization frameworks, multi-modal workflows that incorporate real-world constraints, and studies revealing the potential to inspire creativity, but also highlighting the need for improved user training and interface support.

5.3.1. Summary of Articles on Architectural Design

In [75], Lu et al. present a modular multi-agent framework for generating high-quality, culturally accurate images of traditional Chinese architecture from text prompts. The system comprises five specialized intelligent agents coordinated by a central scheduling agent,

all operating within a closed-loop workflow grounded in a dedicated Chinese Traditional Architecture Cultural Knowledge Base. This structured division of labor enables each agent to focus on distinct tasks. By using the Beijing Central Axis as a case study, comparative experiments demonstrate that the framework significantly outperforms baseline models in image quality, cultural accuracy, and alignment with user intent, while also lowering technical barriers for designers.

In [76], Guida investigates the integration of multi-modal machine learning models into architectural design through three distinct text-to-image-to-3D workflows, demonstrated via a proof-of-concept proposal for the MAXXI Grande Extension competition in Rome. Leveraging CLIP-guided diffusion models such as Stable Diffusion and DALL-E 2, the research explored how designers can engage with AI tools across varying levels of agency: from intuitive, manual prompt-based generation of interior and exterior views to more structured workflows where reconstructed 3D massing models condition image generation. Overall, the study illustrated how multi-modal AI can be flexibly embedded into early-stage architectural workflows to balance automated creativity with contextual precision and fosters trust.

In [77], Paananen et al. explore the role of text-to-image generators in supporting creativity during the initial stages of architectural design, focusing on how these tools influence early-stage ideation and student workflows. The study was conducted via a laboratory workshop with architecture students; it used standardized creativity-support questionnaires, prompt analysis, and group interviews to assess how participants engaged with generative tools. The findings reveal that students approached the tools with diverse creative mindsets and that text-to-image generation can meaningfully enhance conceptual development by fostering discovery and encouraging imaginative thinking. However, the study also found that the effectiveness of generative AI depends strongly on both tool design and user proficiency.

5.3.2. Controllability Requirements of Architectural Design

Across all three papers, controllability is considered essential to balance AI automation with architectural authorship and creative flexibility, though it is managed differently across the workflows.

Lu et al. [75] implement high system-level control through a multi-agent framework—utilizing cultural knowledge bases, LoRA style learning, and automated validation—to minimize user burden and translate vague descriptions into precise, culturally accurate imagery. Conversely, Guida [76] and Paananen et al. [77] argue that control must remain directly with the architect to combat the reductionist simplicity of text prompts. Guida proposes workflows of “varied agency” via prompt tuning and manual image curation, while Paananen et al. demonstrate that out-of-the-box tools offer low built-in control, and thus creative success hinges entirely on user skill, proper selection of constraints, and leveraging prompt ambiguity for discovery.

Ultimately, the papers reject full automation, shifting the definition of control toward iterative, multi-modal workflows that blend structured, automated steering with exploratory flexibility. Users must be able to guide outputs while still benefiting from AI-driven variation and concept expansion.

5.3.3. Domain Constraints in Architectural Design

Domain specificity stands out as the primary hurdle across all studies due to fundamental dataset limitations and a lack of architectural reasoning. Mainstream models like Stable Diffusion or DALL-E 2 are trained on generic internet datasets rather than architecture-specific corpora, creating a “cultural knowledge deficit” regarding spatial

hierarchies and historical symbolism. Although Lu et al. [75] attempt to overcome this limitation by building a specialized archive for traditional Chinese architecture, similar archives would be needed to accommodate other types and variants of architecture.

Because general models lack structured architectural training, they frequently generate visually compelling but structurally implausible forms that ignore spatial logic, material behaviors, floor plans, and climate adaptation. These technical limitations are further compounded by operational and ethical bottlenecks, including high computational costs, long inference times, prompt ambiguity, opaque training sets, embedded cultural biases, and unresolved intellectual property or authorship concerns.

5.3.4. Human-in-the-Loop Considerations for Architectural Design

Rather than replacing professionals, all three papers frame humans as indispensable collaborators whose judgment gives generative AI practical, ethical, and cultural meaning. The authors integrate human expertise in distinct ways, balancing automated system roles with external oversight.

Lu et al. [75] code intent and aesthetic evaluation directly into their multi-agent framework, yet still require external historians and architects to validate outputs for cultural authenticity and appropriate symbolism. Guida [76] externalizes the loop entirely, framing the designer as a critical curator who interacts with intuitive user interfaces to refine datasets, select imagery, and reinterpret forms. Paananen et al. [77] integrate AI directly into an educational studio environment, demonstrating that text-to-image tools work best alongside educator scaffolding, collective critique, and active reflection.

Ultimately, while AI successfully automates early-stage visualization, human evaluation remains the core mechanism required to ensure structural feasibility, stakeholder alignment, and responsible design.

5.4. Fashion Design

Generative-image AI has also been applied to fashion design to help bridge the gap between creative ideation and technical production via tools such as diffusion models and style-transfer systems. The research articles in this category target a diverse set of applications: from deriving production-ready sewing patterns from 2D images, to experimental studio settings where designers engage with tools like Midjourney and Runway to reinterpret historical garments to systems for fashion-focused style transfer offering designers fine-grained control over aesthetic transformations.

5.4.1. Summary of Articles on Fashion Design

In [78], Guo and Sun present a 2D-3D-2D workflow that automates the transformation of garment images into precise 2D sewing patterns. The pipeline begins with the generation of high-fidelity, style-consistent garment images using a DreamBooth-fine-tuned Stable Diffusion model. These images are then reconstructed into detailed 3D models through a Vision Transformer (ViT) encoder and triplane neural radiance fields (NeRFs) to accurately capture structural and textural garment features. The resulting 3D models are processed using surface geometry flattening techniques to produce 2D sewing patterns. Validation through virtual fitting experiments demonstrated the accuracy of the patterns, with dimensional deviations consistently below the industry-accepted threshold of 0.5 cm.

In [79], Rizzi and Bertola investigate the role of generative AI in fashion design through the “Artificial A(i)rchive” laboratory at Politecnico di Milano, Italy, where student teams used tools such as Midjourney, ChatGPT, and Runway to reinterpret Gianfranco Ferré’s iconic 1985 striped jacket. The study explored four modes of human–AI collaboration, Hierarchical, Focused, Casual, and Mutual to investigate how AI can function as a tool for augmented creativity. The results revealed that while generative tools effectively support

ideation and visual exploration, they fall short in capturing the structural complexity, materiality, and bodily relationships essential to garment construction. The findings emphasize that, despite AI's potential to enrich conceptual workflows, the nuanced expertise of fashion designers remains irreplaceable in translating creative vision into wearable form.

In [80], Sun et al. introduce a two-stage neural style transfer model specifically designed to meet the aesthetic and functional needs of the fashion industry. The model enables high-quality, efficient transfer of both global and local style features from inspirational images onto fashion product visuals, while preserving the original garment background. The framework combines Markov-Random-Field-based optimization at reduced resolution with full-resolution Gram-based optimization and incorporates specialized loss functions to support symmetry and gradient effects—key elements in fashion design. The model was shown to provide strong practical utility, allowing designers to quickly and accurately assess how diverse visual inspirations translate onto fashion products.

5.4.2. Controllability Requirements of Fashion Design

Across the three papers, controllability is treated as essential because fashion-generation systems must produce not only visually appealing garments but also technically feasible and manufacturable designs.

Guo and Sun [78] engineer tight structural control by fine-tuning diffusion models on 3D-clothing datasets. They use pose constraints and segmentation guidance to output flattenable 3D mesh patterns with minimal error. Sun et al. [80] argue that large diffusion models act as an unconscious autopilot. They instead use a two-stage, optimization-based style transfer system with explicit mathematical loss functions. Rizzi and Bertola [79] show that practical control is often semantic rather than parametric. Generative models rely on statistical probability instead of a true understanding of fashion meaning; thus, designer control depends on prompt crafting, continuous iteration, and curatorial selection.

Collectively, the papers reject fully autonomous generation; instead, they suggest that effective AI-assisted fashion design requires layered and designer-centered control mechanisms rather than unconstrained generation. Designers need the ability to guide garment structure, preserve brand aesthetics, enforce wearability, and iteratively refine outputs while balancing creativity with technical production constraints.

5.4.3. Domain Constraints in Fashion Design

Across the three papers, recurring barriers include dataset bias and lack of fashion-specific construction knowledge, computational cost of optimization, and the difficulty of evaluating manufacturability versus visual appeal. Mainstream foundation models are trained on generic internet imagery rather than fashion-specific datasets, and thus the employed models frequently generate visually compelling but structurally implausible or conceptually shallow forms.

Interestingly, each paper hits a different fashion-specific wall. Guo and Sun [78] face the geometric challenge of reconstructing deformable 3D garments with folds and pleats; they must ensure these patterns remain flat and suitable for virtual fitting. Sun et al. [80] face aesthetic constraints like true mirrored symmetry and precise background preservation. Rizzi and Bertola [79] point out that industry deployment is restricted by corporate uncertainty regarding intellectual property, originality, and regulatory compliance.

Together, the studies show that fashion AI systems remain constrained not only by technical reconstruction challenges but also by the cultural, semantic, and industrial complexity of professional fashion workflows.

5.4.4. Human-in-the-Loop Considerations for Fashion Design

All three papers strongly frame generative AI as an accelerative assistant rather than a human replacement. The role of the human designer shifts from direct manual drafting toward supervision, curation, and creative direction. AI handles labor-intensive visual exploration and reconstruction, while humans provide functional judgment.

For example, in [78], humans must manually define structural lines on landmarks and rectify darts. In [80], the designer acts as an aesthetic curator who tunes optimization objectives and validates style quality. In the experimental workshops reported by [79], students experimented with image-, text-, and video-based generative AI and then reflected on where these tools expanded design possibilities or weakened design intent, positioning designers as interpreters who supply meaning, enforce technical feasibility, and maintain brand direction.

Collectively, the papers describe a “mutual collaboration” model in which AI handles labor-intensive visual exploration and reconstruction, while humans perform critical tasks: prompt engineering, segmentation correction, virtual fitting checks, ethical oversight, and final creative decision-making.

5.5. Graphic Art

Generative-image AI has also been applied to graphic art by expanding creative control, enhancing stylistic expression, and streamlining complex production workflows for artists, musicians, and designers. The research articles in this category illustrate this application across diverse modalities: from systems that facilitate artistic image transformation to tools that blend lyric-driven generation with mood-based style transfer for album art. Complementary tools provide structured prompt exploration to deepen sense-making in AI art workflows, as well as production-oriented pipelines for game assets.

5.5.1. Summary of Articles on Graphic Art

In [81], Ahn et al. present *DreamStyler*, a one-shot, reference-guided artistic image synthesis framework built on Stable Diffusion that advances both text-to-image generation and style transfer by decoupling style from context while preserving content structure. Unlike traditional methods that often compromise image integrity, *DreamStyler* introduces a specialized denoising pipeline that injects structural conditions from the source image to maintain its form during stylization. Central to its design are two key innovations: Context-Aware Prompt Augmentation, which isolates style features within textual embeddings to enhance stylistic fidelity, and a dual guidance mechanism that separates style and context control, allowing users to independently adjust each component. By optimizing multi-stage textual inversion with these nuanced guidance strategies, *DreamStyler* delivers high-quality, flexible, and faithful artistic transformations across a wide range of reference styles.

In [82], Azuaje et al. present *Visualyre*, which combines lyric-driven image synthesis with mood-based style transfer to generate album artwork. The system derives seed imagery from lyrics, estimates emotional tone from audio, and then blends both signals through a style-transfer stage. In a user study with 35 musicians, more than 60% of participants responded positively, suggesting that this multi-modal pipeline can support accessible music-cover design.

In [83], Ko et al. analyze how visual artists integrate large text-to-image generative models into their workflows. Drawing on a review of 72 papers and interviews with 28 artists, the study identifies three recurring roles for such models: automation, creative exploration, and mediation of ideas and communication. The study also highlights persistent barriers, especially limited control and difficult prompt engineering, and translates them into interface-oriented design guidelines.

In [84], Almeda et al. present *DreamSheets*, a spreadsheet-based interface for systematic exploration of text-to-image design spaces. By embedding LLM-powered prompt functions directly into cells, the system lets users vary semantics, categories, and generation parameters within a structured grid of prompts and outputs. Findings from a lab study and expert deployment indicate that this format supports deliberate, interpretable prompt exploration and stronger creative sense-making.

In [85], Hod et al. present *Playtika*, a system that automates the process of creating resolution-tuned variations of visual assets (graphics) for games. The pipeline decomposes the design process into four key stages: layer extraction from PSD/PSB files, background variation generation, iterative layout refinement, and final harmonization. It leverages LLaVA for generating descriptive prompts, a LoRA model fine-tuned on in-game assets to ensure stylistic consistency, and GPT-4o as an iterative evaluator to optimize object placement. A PIH-based harmonization module seamlessly blends foreground elements with regenerated backgrounds, preserving visual coherence across formats. In a qualitative study involving nine professional designers, eight reported time savings of at least 60%, with overall efficiency gains ranging from 50% to 85%, thus demonstrating the system's potential to streamline asset variation workflows while maintaining the high visual standards.

5.5.2. Controllability Requirements of Graphic Art

Across the five papers, controllability is deemed essential for integrating generative AI into professional creative workflows. Simple text prompting is not enough for creative professionals; instead, the papers collectively argue that creative professionals need interfaces that support iterative, multi-modal, and domain-aware interaction rather than fully autonomous image synthesis. This approach preserves human authority over style and composition, while allowing AI to automate variations and accelerate exploration.

The systems implement control in different ways based on user needs. Visualyre [82] provides a low-threshold interface for musicians using lyric-driven generation and style selection. DreamStyler [81] offers fine-grained diffusion controls by separating artistic style and semantic content via stage-specific embeddings. DreamSheets [84] supports systematic prompt exploration through spreadsheet interfaces that manipulate parameters like seeds and prompt templates at scale. Finally, Playtika [85] provides production-oriented structural control; it integrates game-specific fine-tuning while preserving editable, layered Photoshop files. Ko et al.'s [83] interviews reinforce that artists want varying levels of predictability combined with multi-modal inputs like sketches and brush tools.

5.5.3. Domain Constraints in Graphic Art

The papers indicate that generative AI tools face major technical, artistic, legal, and pipeline constraints in creative industries. Many limitations stem directly from datasets and pretrained models. Across all papers, the core challenge is aligning unpredictable generative models with rigid creative-production standards.

For example, Visualyre [82] relies on small, generic datasets, which hinders its ability to represent abstract concepts or non-English lyrics. DreamStyler [81] struggles to disentangle highly abstract styles and faces risks of memorizing copyrighted art. Evolving model behavior, opacity, and computational latency also create unpredictability for users. Production workflows also introduce domain-specific engineering constraints. Playtika [85] must handle multiple aspect ratios and resolutions and integrate into existing design pipelines, all while maintaining stylistic consistency. DreamSheets [84] exposes additional limitations involving computational latency, opacity of prompt-to-image mappings, and rapidly evolving model behavior that artists must learn through experimentation. Ko et al.'s study [83]

highlights broader industry concerns such as copyright infringement and unlicensed training data.

Across all five papers, recurring constraints include unpredictability, computational cost, stylistic inconsistency, insufficient domain-specific knowledge, and unresolved ethical and copyright concerns surrounding professional use of generative AI.

5.5.4. Human-in-the-Loop Considerations for Graphic Art

All five papers frame AI as a co-creative assistant that augments rather than replaces human judgment. Humans remain responsible for defining intent, evaluating outputs, and making final design decisions; meanwhile, the AI handles brute-force variation and reduces repetitive, mechanical labor. The designer's role shifts from direct manual production toward orchestration, curation, and creative supervision.

The level of human involvement varies across the workflows. Visualyre [82] positions users as high-level curators who select preferred style options. DreamStyler [81] requires quick manual caption editing to fix style-content entanglement. DreamSheets [84] treats prompting as an exploratory foraging process where users scan hundreds of variants to build a mental model of the AI's behavior. Playtika [85] automates massive resizing workloads but intentionally returns fully editable layers so designers retain final creative freedom. Ko et al. [83] note that artists value AI for automation, exploration, and client communication, but they must always lead the process.

5.6. Storytelling

Storytelling can benefit substantially from effective visuals, making it a natural application area for generative-image AI. The research articles in this category reflect that trend: from narrative-driven platforms that translate textual intent into emotionally resonant, visually consistent story imagery, to tools that integrate diffusion models and Gaussian Splatting for filmmakers. In design-oriented contexts, systems have also been developed to streamline the creation of coherent storyboards using automated scene segmentation, sketch guidance, and textual prompting. Together, these studies underscore persistent challenges in alignment, consistency, and user control.

5.6.1. Summary of Articles on Storytelling

In [86], Leininger et al. report a pilot study on the practical value of three AI-generated 3D environment types in filmmaking workflows: depth meshes, panoramic meshes, and Gaussian Splatting. Using a prototype called *EnVisualAIzer* and feedback from 15 industry experts, the study shows that different representations support different tasks, from early ideation to high-fidelity previsualization. The main contribution is an empirical comparison grounded in professional practice rather than purely technical evaluation.

In [87], Fernandes et al. present *ArtAI4DS*, a digital-storytelling tool that converts story-derived keywords into images intended to support creative expression and emotional engagement. The system emerged from a Wizard-of-Oz phase and subsequent co-design refinement, with YAKE selected as the most suitable keyword extractor for guiding SDXL generation. Final evaluations with seven participants suggested that the interface supports creativity and social interaction, but also revealed a need for clearer onboarding and more precise keyword guidance.

In [88], Shahriyar et al. present *Bibliosmia*, a framework for generating personalized children's stories with consistent text-image alignment. Its three-module pipeline combines GPT-4o-based narrative generation, structured alignment, and identity-consistent image synthesis driven by a multi-modal diffusion transformer. Evaluated on an online storytelling platform, the system achieved strong prompt similarity and character consistency results while reducing the need for manual regeneration.

In [89], Liang et al. present *StoryDiffusion*, which combines GPT-4 and Stable Diffusion to automate scene segmentation, stylistic prompt construction, and storyboard rendering. The system is intended to keep visual sequences stylistically coherent while preserving iterative control over both text and images. A user study with 12 UX design students reported time savings, lower workload, and improved storytelling quality relative to hand-drawing.

In [90], Chan et al. present *SketchBoard*, a storyboarding system that combines sketch guidance, text prompting, and LLM-based narrative support. Built around Stable Diffusion, ControlNet, and GPT-3.5, the system translates rough sketches and prompts into coordinated visual and textual story sequences. A user study with 50 students reported high acceptance, while also identifying occasional text–image mismatches as a remaining limitation.

5.6.2. Controllability Requirements of Storytelling

Across the five papers, controllability is treated as essential for making generative AI useful in storytelling, storyboarding, and virtual-production workflows. Rather than relying on pixel-level editing, these systems shift control toward higher-level semantic inputs such as as natural-language prompts, sketches, and character profiles.

StoryDiffusion [89] lets users iteratively regenerate storyboard frames from narrative descriptions using two modes: exploratory AI-directed drafts and user-directed refinement. SketchBoard [90] adds spatial control by using rough sketches with ControlNet and Stable Diffusion to preserve storyboard layouts and game-specific styles. Bibliosmia [88] demands structured control through child profiles, persistent character sheets, and running state tracking to enforce identity consistency.

Control mechanisms also vary based on domain and user expertise. ArtAI4DS [87] prioritizes emotional control for non-artists through editable prompts, revisitable generations, and keyword transparency. EnVisualAIzer [86] shifts control entirely to spatial navigation and real-time scene manipulation inside Unreal Engine 5 to manipulate, e.g., lighting, camera placement, and object positioning.

Together, the papers argue that generative storytelling systems must support flexible, iterative, and context-sensitive control mechanisms that preserve narrative intent, stylistic coherence, and production usability.

5.6.3. Domain Constraints in Storytelling

Across all five papers, recurring constraints include bias and copyright concerns inherited from pretrained models, dependence on large external APIs and datasets, computational costs, stylistic inconsistency, limited contextual reasoning, and difficulties aligning generated content with the nuanced emotional and cultural expectations of creative storytelling domains.

Bibliosmia [88] faces particular strict constraints because it targets children’s education and social-emotional learning. Stories must be age-appropriate, culturally inclusive, and emotionally coherent, requiring vision–language model validation to filter out inappropriate content. StoryDiffusion [89] and SketchBoard [90] similarly struggle with maintaining visual and narrative continuity over multi-scene storyboards, especially when integrating text generation with image synthesis while preserving recognizable characters and consistent style.

EnVisualAIzer [86] highlights challenges related to integrating AI-generated environments into real-time filmmaking pipelines, including navigability, interoperability with tools such as Unreal Engine, and balancing fidelity against interactivity. ArtAI4DS [87] raises concerns about whether AI-generated visuals accurately preserve autobiographical meaning and emotional nuance for users sharing personal migration stories.

5.6.4. Human-in-the-Loop Considerations for Storytelling

All five papers strongly position AI as a co-creative assistant or creativity support co-pilot rather than a human replacement. Human users remain responsible for defining narrative goals, emotional tone, and contextual meaning. Meanwhile, the AI automates labor-intensive processes like layout generation, style consistency, and multi-view reconstruction. Creative value emerges from an iterative loop where the model expands the possibility space and the human orchestrates. The human role clearly shifts from manual asset creation toward curation, evaluation, and refinement; and ultimately, artistic judgment and final validation remain firmly under human supervision.

The exact structure of the human loop varies. StoryDiffusion [89] and SketchBoard [90] cast the designer as an orchestrator who submits sketches, reviews variations, and resubmits prompts. ArtAI4DS [87] utilizes participatory co-design workshops where users compare hand-drawn art to AI outputs to protect personal authorship. Bibliosmia [88] formalizes oversight at both ends by having caregivers define child profiles upfront and review the final assembly for developmental appropriateness. EnVisualAIzer [86] supports role-specific loops where directors use panoramas for rapid world-building and cinematographers use splats for lighting previsualization.

5.7. Three-Dimensional Scene Generation

Generative-image AI has been applied to 3D scene generation to enable workflows that blend real-world context, object-level control, and photorealistic synthesis from minimal user input. As illustrated by the research articles in this category, the general trend involves combining real-world data with diffusion models to create coherent, navigable, and photorealistic virtual environments from limited input. Recent work emphasises modularity and spatial consistency, using techniques such as independent object editing via NeRFs and progressive scene expansion guided by depth optimization. By refining single images or fragmented captures into high-fidelity 3D scenes, these technologies improve the accessibility and controllability of immersive 3D environment creation.

5.7.1. Summary of Articles on 3D Scene Generation

In [91], Numan et al. present *SpaceBlender*, a framework for creating VR telepresence environments by blending reconstructed real-world submeshes with language-guided diffusion inpainting. The two-stage pipeline joins local reconstructions into a navigable virtual scene and then fills transitional gaps to improve coherence. A user study with 20 participants reported gains in comfort and navigability relative to baseline text-to-3D systems, although geometric fidelity remained a limitation.

In [92], Epstein et al. present an unsupervised text-to-3D framework that produces scenes composed of disentangled, independently controllable objects. Its central idea, *Layout Learning*, optimizes multiple NeRFs across varied spatial arrangements so that each learns a coherent object representation. Compared with more monolithic scene generators, this design improves editability and object-level control without requiring auxiliary supervision.

In [93], Zhang et al. present *Text2NeRF*, a framework that combines text-to-image diffusion with NeRF optimization to build geometrically consistent 3D environments from a single prompt. The method expands scenes progressively through depth-guided inpainting and multi-view updates, allowing geometry and appearance to co-evolve across viewpoints. Its main contribution is to show how diffusion priors can be coupled with explicit 3D constraints to improve realism and spatial coherence.

In [94], Pu et al. present *Pano2Room*, a framework for reconstructing indoor 3D scenes and synthesizing novel views from a single panorama. The pipeline refines a coarse back-

projected mesh with diffusion-guided texture and depth estimation, then converts the result into a 3D Gaussian Splatting representation. Across multiple datasets, the method outperformed earlier approaches, indicating the value of combining panoramic geometry with scene-specific diffusion refinement.

5.7.2. Controllability Requirements of 3D Scene Generation

Across the four papers, controllability is treated as essential for making AI-generated 3D and VR environments practically usable, editable, and trustworthy in creative and collaborative workflows. However, the degree of controllability offered by the various systems spans a rather large spectrum.

Text2NeRF [93] and Pano2Room [94] sit at the lowest-touch end: Text2NeRF needs only a natural-language prompt, then uses monocular depth priors and a progressive inpainting-and-updating loop to grow a NeRF view-by-view while enforcing multi-view consistency. Pano2Room similarly requires just one phone-captured panorama, then iteratively inpaints occlusions and bakes a 3D Gaussian Splatting field.

Layout Learning [92] and SpaceBlender [91] add higher-level structural control. Layout Learning stays text-only at input but jointly optimizes multiple disentangled NeRFs and learned layouts, defining objects as independently movable parts; thus, users can reposition, freeze, or recombine objects without manual annotation. SpaceBlender is the most user-grounded: participants supply photos of their physical spaces, and the pipeline blends scenes, with explicit controls over privacy settings, realism, blending intensity, and inter-spatial distance.

Collectively, the papers converge on the idea that controllability in generative 3D systems must operate at multiple levels—including object manipulation, geometry preservation, spatial coherence, realism, and user steering—while minimizing tedious manual configuration and preserving usability for non-expert users.

5.7.3. Domain Constraints in 3D Scene Generation

Across all four papers, recurring constraints include the scarcity of high-quality 3D training data, limited scene grounding (linking text-based descriptions to specific 2D/3D objects), inconsistent geometry, weak scalability, and difficulties maintaining realism and consistency in complex spatial environments. Across all systems, computational cost and long optimization times also remain significant bottlenecks, particularly for NeRF- and diffusion-based pipelines.

Text2NeRF [93] and Layout Learning [92] both rely heavily on pretrained 2D diffusion models because large-scale paired text-and-3D datasets remain limited. This reliance introduces issues such as distorted geometry, weak object boundaries, unstable view consistency, and “dreamlike” scene artifacts. Pano2Room [94] similarly addresses the difficulty of reconstructing complete indoor scenes from sparse panoramic input, especially in the presence of occlusions, reflective surfaces, or cluttered environments. SpaceBlender [91] highlights additional constraints tied to VR usability, including navigability, user comfort, privacy, and social interaction quality within generated spaces.

The papers also point to broader ethical and practical concerns, including dependence on internet-scale pretrained models, inherited dataset biases, and potential privacy risks when users upload images of personal environments for immersive scene generation.

5.7.4. Human-in-the-Loop Considerations for 3D Scene Generation

All four papers position humans as creative supervisors, evaluators, and collaborators rather than passive consumers. The systems automate low-level 3D reconstruction and optimization. This minimizes the technical burden on the user. Humans supply high-value

context like prompts, panoramas, or room photos. They define the semantic intent, while the AI manages the brute-force rendering and scene completion.

The exact interaction workflow varies across the systems. Text2NeRF [93] and Pano2Room [94] focus on zero-shot generation followed by a “generate then inspect” loop; thus, human judgment appears during the final qualitative evaluation of realism and artifacts. Layout Learning [92] and SpaceBlender [91] offer greater user interaction/control, but also require increased technical burden. Together, the papers portray human-in-the-loop interaction as fundamental; human inspection remains the core mechanism to ensure factors such as realism, creative intent, and contextual appropriateness.

5.8. Three-Dimensional Character Generation

Generative-image AI has also been applied to 3D character creation to enable the generation of high-quality controllable avatars from single images or text prompts. As illustrated by the research articles in this category, the reviewed approaches include pose-canonicalization pipelines, frameworks that disentangle bodies and garments using compositional NeRFs, methods that combine skeleton-guided diffusion with domain-specific losses and multi-view consistency techniques, and systems designed to perform semantic decomposition to enhance geometric cleanliness and downstream usability.

5.8.1. Summary of Articles on 3D Character Generation

In [95], the authors present *CharacterGen*, an end-to-end framework for reconstructing 3D characters in a canonical A-pose from a single 2D image. The system first synthesizes consistent multi-view imagery and then performs coarse-to-fine 3D reconstruction with texture back-projection. A user study with 21 participants indicated a clear preference for its outputs over earlier methods.

In [96], the authors present *TELA*, a text-to-3D framework that separates body and garments into distinct NeRF components. Through stratified compositional rendering and dual score-distillation losses, the method produces more controllable and structurally disentangled human models than competing approaches. This separation is particularly relevant for downstream tasks such as virtual try-on and garment transfer.

In [97], the authors present *AniDream*, a framework for anime-style 3D avatar generation. It combines skeleton-guided score distillation, anime-specific normalization and inpainting losses, and ControlNet-based prompt guidance to balance stylistic fidelity with anatomical coherence. Quantitative evaluation suggests gains in semantic alignment, stylistic consistency, and visual quality.

In [98], the authors present *SeparateGen*, a framework that reconstructs semantically decomposed 3D characters from a single posed image. By separating body, clothing, hair, and shoes and reconstructing them with component-aware decoders, the method reduces mesh fusion artefacts and improves reuse potential. The reported gains concern both visual quality and cross-view consistency.

5.8.2. Controllability Requirements of 3D Character Generation

Across the four papers, controllability is treated as a core requirement for practical AI-based 3D character and avatar generation, particularly for animation, rigging, virtual try-on, and AR/VR deployment workflows. All four systems lower the entry barrier to 3D character creation by replacing manual modeling with simple cues, but they distribute control differently across the pipeline.

CharacterGen [95] requires the least user input: one arbitrary-pose image is automatically converted into a canonical animation-ready 3D character through pose normalization and sparse-view reconstruction. SeparateGen [98] extends this pipeline by semantically separating the body, clothing, hair, and shoes into independently editable components.

TELA [96] and AniDream [97] shift control to language and structure. TELA gives layer-wise, text-driven control while enabling virtual try-on and garment replacement. AniDream focuses on user-friendly textual and skeletal control mechanisms for anime avatar generation, allowing users to specify appearance details via prompts.

Collectively, the papers converge on the idea that usable character-generation systems can benefit from controllability at multiple levels, while minimizing the amount of manual technical intervention required by users. These opposing objectives are met by shifting controllability from vertex-level manipulation to semantic/style-level manipulations.

5.8.3. Domain Constraints in 3D Character Generation

A major shared constraint across the papers is the lack of large-scale datasets suitable for stylized or anime-style 3D character generation. Existing reconstruction datasets and body priors are largely designed for realistic humans and struggle with exaggerated anatomy, layered garments, nonstandard poses, and stylized textures.

CharacterGen [95] and SeparateGen [98] address this limitation by introducing custom datasets (the Anime3D and SC-Anime datasets, respectively). TELA [96] addresses the absence of layered text-to-3D clothing datasets by relying on pretrained 2D diffusion priors. AniDream [97] compensates for limited anime training data using lightweight LoRA adaptation and geometry regularization.

The systems also face broader technical and practical constraints, including sparse-view ambiguity, self-occlusion, mesh entanglement, artifacts/inconsistencies, and high computational cost. In addition, several papers implicitly raise concerns about the legal and ethical implications of using scraped anime assets and publicly sourced character data for training and deployment.

5.8.4. Human-in-the-Loop Considerations for 3D Character Generation

Traditional 3D character creation requires significant technical expertise and manual labor, but these systems automate difficult tasks such as pose correction, reconstruction, semantic decomposition, and stylization. In this way, all four papers frame AI as a co-creative tool that assists rather than replaces artists and animators.

Overall, the papers aim to reposition humans as directors, curators, and downstream integrators. CharacterGen [95] minimizes modeling effort by generating animation-ready meshes, while SeparateGen [98] and TELA [96] mirror professional workflows by producing modular, editable character components that support rigging, clothing transfer, and iterative customization. AniDream [97] similarly enables users to rapidly create anime avatars through simple text prompts without requiring advanced modeling skills.

Nonetheless, human input remains essential throughout the workflow. Users still provide creative direction through image selection, text prompting, editing decisions, semantic data annotation, and perceptual evaluation. In this way, AI handles technical generation tasks while humans retain artistic control and interpretive judgment.

5.9. Open Challenges

Cross-cutting analysis of the 29 retained application studies reveals a set of technically grounded challenges that persist across domains and that reduce the transition from research prototype to production deployment.

Deployment maturity and output reliability. The majority of the reviewed studies operate at a research prototype or proof-of-concept stage: evaluation cohorts are small (commonly 7–35 participants), deployment durations are short, and success criteria are often task-specific rather than production-grade. This maturity gap is most visible in the domain-critical applications: the culturally grounded architectural visualization framework of [75] demonstrates strong benchmark results within a closed knowledge base but does not address out-of-

distribution prompts or robustness to adversarial inputs at scale; the 2D-to-3D fashion pipeline of [78] achieves sub-0.5 cm dimensional accuracy in controlled garment categories, yet the required DreamBooth fine-tuning per garment style imposes a per-deployment computational overhead that has not been evaluated at production volume. The evidence base therefore supports the conclusion that domain-specific accuracy is achievable in controlled conditions, but that scalability, latency, and failure-mode cataloguing remain largely unaddressed.

Controllability constraints and spatial fidelity. A recurring technical limitation in the application literature is the weakening of spatial precision as generation scales to complex multi-object scenes or cross-view consistency requirements. Studies in storytelling [88,89] and graphic art [85] each report residual inconsistency in multi-panel or multi-view outputs, where identity or style information drifts across scenes without explicit inter-frame constraints. The pattern is consistent with the broader Janus and semantic-drift problems documented in the generative architecture literature for 3D and panoramic synthesis—indicating that application-layer composability requirements expose the same underlying architectural limitation. Prompt sensitivity is a related constraint: small prompt perturbations induce large output variance, and several user studies report that non-expert practitioners require extensive prompt iteration to achieve acceptable results [77,83], suggesting that current text-conditioning interfaces are not yet aligned with domain practitioner workflows.

Human-in-the-loop design and augmentation versus over-reliance. Across design, fashion, and storytelling applications, the studies consistently reveal a bifurcated user response to generative tools: when embedded in structured workflows with clear expert roles (e.g., the multi-agent architecture of [75], the structured storyboard pipeline of [89]), AI augments output quality and reduces iteration time. When the interface exposes the full generative distribution without workflow scaffolding, users report creative hindrance, loss of ownership, or over-reliance on generated suggestions at the expense of original intent [77,79]. This finding has direct implications for interface design: the effective deployment of generative image AI in professional settings likely requires the explicit modeling of the human decision boundary—distinguishing which stages should remain human-controlled, which should be AI-augmented, and which can be automated—rather than offering unstructured access to the generative capacity of the underlying model. The human-in-the-loop pattern documented in the film pre-visualization study [86] and the spreadsheet-based prompt exploration system [84] provides partial evidence that structured interfaces can preserve user agency while still realising efficiency gains, but formal comparative evaluation across domains remains absent from the current literature.

- *Emotional Influence in Co-Creation:* AI-generated visuals can evoke strong emotional responses that can unintentionally sway user perceptions and design decisions. How to design improved systems and how to better train designers to maintain objectivity and creative clarity remains an open research challenge.
- *Interpretability and Transparency:* The black-box nature of many generative models limits their integration into professional workflows. Future research is needed on optimal ways to offer structured reasoning and clearer insight into how outputs are generated.
- *Prompt Engineering and Interface Design:* Artists and designers often struggle with complex prompt creation and unintuitive tools. Streamlining these processes and offering domain-specific customization remains a challenge.
- *Aligning Visuals with Narrative Intent:* Bridging the gap between user intent and AI output remains a challenge, especially in storytelling and feedback contexts. Ensuring that generated visuals accurately reflect symbolic meaning and narrative flow has proved difficult.

- *Maintaining Consistency Across Views/Time*: Achieving visual, stylistic, and identity consistency across multiple scenes, angles, or temporal sequences remains challenging, particularly in character generation and storyboard development.
- *Ensuring Contextual and Cultural Fidelity*: Generating content that respects cultural nuances, architectural traditions, or real-world constraints requires integrating specialized knowledge bases and improving the contextual awareness of generative systems.

5.10. Section-Specific Limitations

The applications corpus combines studies with substantially different maturity levels, from exploratory prototypes to domain-embedded pilot deployments, and often relies on relatively small user cohorts and context-specific evaluation settings. Because tasks, success criteria, and reporting depth differ across design, storytelling, and 3D scenarios, direct quantitative comparability is limited. The synthesis in this section therefore emphasises recurring design patterns, deployment constraints, and reproducible methodological signals rather than universal performance ordering.

6. Multi-Modal and Cross-Lingual Aspects of AI-Generated Scientific Content

This section surveys research that sits at the intersection of multi-modality and language diversity in the context of AI-generated scientific content, i.e., systems that jointly model text and visual signals, transfer meaning across modalities, and remain usable when prompts, captions, or downstream communication occur in multiple languages. Beyond “better generation”, the core theme here is the controllable and interpretable interaction between modalities: how models align representations, ground language in visual evidence, and move information reliably across text–image (and sometimes video) channels, including under cross-lingual conditions. The discussion is organized around seven recurring directions, spanning unified foundation architectures and generative multi-modal models, representation learning and cross-modal alignment, text-to-image and image-to-text pipelines, temporal/video extensions, behavioral effects and bias, and multi-lingual transfer mechanisms.

6.1. Section-Specific Selection and Thematic Grouping

Within the global review workflow described in Section 2, the multi-modal and cross-lingual branch was subjected to an additional section-specific screening procedure. Starting from the 306 records assigned to this thematic category during the rule-based assignment stage, a two-step filtering process was applied to isolate studies strictly aligned with multi-modal generative modeling and cross-lingual or cross-modal content generation. In the first step, domain-specific and generative constraints were enforced to remove clearly irrelevant multi-modal works, such as studies focused purely on classification, sentiment analysis, robotics, or perception-only tasks, reducing the pool to 74 candidate records. In the second step, a stricter semantic consistency review was conducted to retain only papers explicitly addressing multi-modal generation, cross-modal alignment, multi-lingual or cross-lingual transfer, or unified multi-modal foundation architectures relevant to AI-generated scientific content. This yielded a final core corpus of 31 articles used in the present section.

Two borderline studies explicitly flagged during refinement (Lin et al. [99] and Ramu et al. [100]) were retained because their central contributions target fine-grained cross-modal grounding and region-sensitive image-to-text semantic transfer, which directly support the section objective of analysing multi-modal alignment mechanisms rather than standalone unimodal captioning performance.

The final set of 31 core articles was subsequently organized into thematic clusters reflecting the dominant research directions within multi-modal and cross-lingual AI-generated content. Clustering was performed through qualitative inspection of each work’s primary objective, modeling strategy, and modality interaction pattern. Seven coherent clusters emerged, capturing foundation and unified architectures (Section 6.2), representation and alignment learning (Section 6.3), text-to-image generation (Section 6.4), image-to-text generation (Section 6.5), video and temporal modeling (Section 6.6), bias and behavioral analysis (Section 6.7), and multi-lingual or cross-lingual transfer mechanisms (Section 6.8). This structured grouping enables a more systematic narrative synthesis while preserving the conceptual distinctions between architectural, generative, and analytical contributions. Table 4 presents the retained study characteristics, and the following subsections synthesize findings within these clusters.

The following subsections provide a detailed synthesis of the findings from the selected articles, organized according to the identified thematic categories.

Table 4. Summary of representative works on multi-modal and cross-lingual aspects of AI-generated scientific content (Part I) & (Part II).

Part I				
Reference	Year	Modalities/Data	Methodological Core	Key Contribution
<i>Foundation and Unified Architectures</i>				
Cho et al. [101]	2020	Text + Image	Introduces X-L XMERT, an extension to LXMERT with training refinements including discretizing visual representations.	Connects enhanced multi-modal pretraining to stronger generation-oriented transfer (e.g., captioning-style tasks).
Yu et al. [102]	2023	Text + Image + Video	Introduces Semantic Pyramid Auto Encoder (SPA-E) to enable frozen LLMs to handle multi-modal understanding and generation.	Validates in-context multi-modal learning with frozen LLM backbones across diverse tasks.
Fang et al. [103]	2024	Text + Image + Video + Audio	Proposes i-Code Studio, a configurable and composable framework for integrative multi-model systems.	Positions integrative, multi-model composition as a practical route toward broader multi-modal capability.
Jain et al. [104]	2024	Text + Image	Proposes a “versatile code” style design for improved multi-modal perception and reasoning in MLLMs.	Targets more reliable vision–language reasoning via structured intermediate representations.
Pan et al. [105]	2024	Text + Image	Proposes an auto-encoding style framework for multi-modal generation with shared representation learning.	Demonstrates a unified encoding/decoding pathway for multi-modal content synthesis.
Tang et al. [106]	2024	Text + Image + Video	Presents CoDi2-style unified generation across multiple modalities via coordinated token spaces.	Improves any-to-any generation flexibility while keeping a coherent cross-modal interface.
Huang et al. [107]	2025	Text + Image	Proposes an autonomous multi-modal agent pipeline that plans and executes generation/editing via tool use.	Shows end-to-end, tool-augmented multi-modal creation workflows with tighter control loops.
Cui et al. [108]	2026	Text + Image	Presents models as in-context learners, emphasizing instruction-conditioned generalization.	Highlights in-context multi-modal learning as a unifying capability for both understanding and generation.
<i>Representation Learning and Cross-Modal Alignment</i>				
Wu et al. [109]	2022	Audio + Image + Text	Aligns audio to a CLIP-style joint embedding space for cross-modal retrieval and transfer.	Enables audio–vision–language interoperability by mapping audio into vision–language representations.
Lin et al. [99]	2023	Text + Image	Proposes a cross-modal matching/alignment strategy to improve consistency across modalities.	Strengthens cross-modal retrieval/grounding by tightening shared embedding geometry.
Yang et al. [110]	2023	Text + Image	Introduces semantic alignment constraints to better couple language and visual representations.	Improves semantic consistency between modalities, supporting downstream generative/grounded tasks.
Chen et al. [111]	2024	Text + Image (large-scale data)	Releases and curates a large-scale paired dataset and training protocol to support robust multi-modal alignment.	Provides scale and curation signals that improve alignment quality for training multi-modal models.
Fang et al. [112]	2024	Text + Image	Explores representation and interaction mechanisms that affect cross-modal alignment behavior in practice.	Identifies alignment-sensitive design factors that influence robustness and semantic fidelity.

Table 4. Cont.

Part I				
Reference	Year	Modalities/Data	Methodological Core	Key Contribution
Bansal et al. [113]	2024	Video + Text	Proposes a video–text contrastive learning setup to improve temporal cross-modal correspondence.	Enhances temporal grounding by enforcing stronger video–language alignment objectives.
Campi et al. [114]	2025	Text + Image	Proposes and assesses alignment objectives aimed at more faithful cross-modal semantic correspondence.	Clarifies practical trade-offs between alignment strength and task-specific performance.
Part II				
Reference	Year	Modalities/Data	Methodological Core	Key Contribution
<i>Text-to-Image Generation</i>				
Hutchinson et al. [115]	2022	Text→Image	Analyzes underspecification and ambiguity in text prompts for image generation.	Shows how prompt ambiguity drives variability and motivates more explicit controls/evaluation.
Wang et al. [116]	2024	Text + Image guidance	Proposes multi-modality-guided generation strategies to better steer synthesis.	Improves controllability/faithfulness by injecting additional modality cues into generation.
Abdel et al. [117]	2025	Text→Image	Studies whether and how multi-modal constraints improve text-to-image generation behavior.	Provides evidence on when multi-modal conditioning helps (and when it can fail) for faithful synthesis.
Yang et al. [118]	2025	Text→Image (ideation)	Proposes an idea-to-image workflow linking structured intent representations to image generation.	Supports more systematic translation from high-level intent to visual outputs.
Hong et al. [119]	2026	Text→3D/Avatar	Presents text-conditioned generation for 3D avatars with animation as an output target.	Extends text-conditioned generation beyond 2D imagery to controllable 3D + motion content.
<i>Image-to-Text Generation</i>				
Tewel et al. [120]	2022	Image→Text	Propose zero-shot image captioning via CLIP-like alignment and language modeling.	Enables captioning without paired caption training by leveraging vision–language priors.
Elaszczyk et al. [121]	2023	Image + Text	Propose a cross-modal setup for improved image-to-text generation/grounding.	Strengthens captioning/description fidelity by tighter cross-modal coupling.
Ramu et al. [100]	2024	Image→Text (local details)	Propose “zooming”/region-focused captioning to capture fine-grained visual evidence.	Improves descriptive specificity by explicitly attending to local visual content.
Fernandes et al. [122]	2025	Image→Text (domain eval)	Propose a dataset/benchmarking protocol for image description assessment in a targeted setting.	Provides evaluation structure for comparing captioning systems under domain constraints.
<i>Video and Temporal Generation</i>				
Luo et al. [123]	2024	Music/Audio + Video	Propose audio-/music-conditioned generation for temporally coherent dance/motion video.	Demonstrates cross-modal temporal control where audio structure steers generated motion/video.
Wu et al. [124]	2024	Image + Text→Video/Animation	Propose an animation pipeline that leverages multi-modal conditioning for temporal synthesis.	Improves temporal consistency/controllability in multi-modal animation generation.
<i>Bias, behavior, and Emergent Properties</i>				
Chen et al. [125]	2024	multi-modal LMs (behavior)	Analyze behavioral properties of multi-modal models under controlled prompting/evaluation.	Surfaces failure modes and emergent behaviors relevant to reliable scientific use.
Aman et al. [126]	2025	Text + Image (bias probe)	Propose/compile probes that reveal bias or systematic behaviors in multi-modal systems.	Provides targeted evidence of bias-like effects and motivates mitigation/evaluation practice.
<i>multi-lingual and Cross-Lingual Transfer</i>				
Long et al. [127]	2021	Text (multi-lingual) + Vision	Propose multi-lingual generation/transfer mechanisms relevant to multi-modal settings.	Demonstrates cross-lingual generalization as a core requirement for accessible multi-modal generation.
Mohammed et al. [128]	2022	Multi-lingual Text + Vision	Propose bootstrapping strategies to extend multi-modal capability to additional languages.	Shows weak supervision/bootstrapping expands coverage beyond resource-rich languages.
Onami et al. [129]	2024	Doc + Image + Text (Japanese)	Present a document QA setting supporting Japanese and cross-lingual multi-modal understanding.	Highlights document-centric multi-lingual evaluation needs for practical scientific workflows.

6.2. Foundation and Unified Multi-Modal Architectures

Early attempts to extend large language models to multi-modal settings focused primarily on strengthening cross-modal understanding while maintaining a text-centric backbone. Cho et al. [101] introduced a multi-lingual extension of vision language pre-training, demonstrating that shared transformer encoders can align visual regions with linguistic representations across languages. Although generation was not the primary objective, this line of work established the importance of unified token spaces and cross-modal pre-training as a foundation for more integrated architectures.

A complementary direction explores how frozen or minimally adapted LLM can be repurposed for multi-modal generation. Yu et al. [102] proposed the Semantic Pyramid Auto Encoder (SPAEE), which maps images to interpretable lexical tokens derived from an LLM vocabulary, allowing frozen language models to understand and generate visual content through in-context learning. By organizing tokens in a semantic pyramid structure, SPAEE bridges discrete linguistic representations with pixel-level reconstruction while preserving compatibility with general-purpose LLMs. Similarly, Pan et al. [105] introduced morph-tokens to decouple visual abstraction (for comprehension) from visual completeness (for generation). Their three-stage training strategy resolves the inherent tension between understanding and reconstruction within autoregressive multi-modal LLMs, demonstrating that unified token-based architectures can simultaneously support captioning, editing, and image synthesis.

More recent efforts emphasize scalability and autonomy in unified multi-modal systems. Fang et al. [103] advanced the notion of integrated multi-modal coding by aligning vision representations more closely with LLM token spaces, strengthening the bidirectional flow between perception and generation. Huang et al. [107] moved further towards autonomous multi-modal agents, where unified architectures are designed not only for conditional generation but also for goal-driven multi-modal reasoning and iterative refinement. Finally, Cui et al. [108] investigated multi-modal foundations of generative-first, arguing for architectures that treat multi-modal generation, not only understanding, as a primary training objective, thus reshaping the design of multi-modal unified transformers.

Taken together, these works illustrate a clear evolution: from shared encoder-based alignment models through tokenization strategies that translate visual content into linguistic spaces to fully unified architectures where a central LLM orchestrates multi-modal perception and generation. The emerging consensus is that effective multi-modal foundations require both representational alignment and architectural decoupling mechanisms that reconcile abstraction with faithful reconstruction, enabling scalable any-to-any multi-modal generation within a single coherent framework.

6.3. Representation Learning and Cross-Modal Alignment

Learning unified representations across modalities is central to modern multi-modal systems, particularly in image-text and audio-visual retrieval, generation, and zero-shot transfer. A dominant paradigm is contrastive pre-training, where modality-specific encoders are trained to project inputs into a shared embedding space such that semantically aligned pairs are close while mismatched pairs are separated. Representative examples include CLIP-style dual-encoder frameworks and their extensions to additional modalities.

In the audio-visual domain, distillation-based approaches have proven effective. Wav2CLIP projects audio into the CLIP image-text embedding space by freezing the visual encoder and training an audio encoder to predict CLIP image embeddings from video audio streams [109]. This strategy avoids re-learning the visual model and yields audio representations aligned with both images and text, enabling zero-shot classification and cross-modal retrieval. Similarly, WavBriVL extends this idea by distilling from BriVL,

a large-scale vision–language model trained with a memory-bank-based contrastive objective, and aligns audio with image and text embeddings in a shared space [112]. The broader trend of using pre-trained multi-modal backbones for efficient adaptation is also reflected in recent large-scale multi-modal alignment efforts [111].

Beyond coarse-grained global alignment, recent work emphasizes fine-grained region–word correspondence. Although dual-encoder contrastive models operate primarily at the global embedding level, fine-grained methods attempt to align local image regions with textual tokens to reduce modality gaps and improve retrieval accuracy. For example, semantic completion and filtration mechanisms explicitly address incomplete textual descriptions by generating enriched semantic representations and filtering irrelevant region–word pairs [110]. Cross-modal graph-based matching frameworks further refine alignment by modeling structural dependencies between regions and words [99].

Recent multi-modal video-text alignment approaches also highlight the importance of temporal modeling and structured cross-modal aggregation for representation learning [113]. Moreover, unified contrastive frameworks that extend alignment across additional modalities or tasks demonstrate that scalable pre-training objectives can serve as a strong foundation for downstream multi-modal retrieval and generation systems [114].

In general, cross-modal representation learning hinges on two complementary principles: (1) enforcing shared semantic geometry across modalities through contrastive or distillation objectives and (2) refining alignment at finer semantic granularity to suppress spurious correspondences and capture structural relationships. Integrating these perspectives provides a principled foundation for robust multi-modal retrieval and generation systems.

6.4. Text-to-Image Generation

Text-to-image generation has matured from a collection of narrowly scoped modeling strategies into a broader ecosystem of multi-modal generative systems, evaluation frameworks, and domain-specific workflows. A useful starting point for framing this shift is the observation that scene description-to-depiction tasks are inherently underspecified: the same textual description can legitimately admit multiple visual realizations, and the ambiguity is not a corner case but a defining property of the problem [115]. This perspective matters in practice because it directly affects what it means to “evaluate” a model, what should be treated as an error versus a valid interpretation, and where ethical risks may be amplified by implicit choices made during visualization [115].

With the growing real-world impact of T2I systems, an adjacent line of work focuses on the auditing and interpretability of prompt–image associations. Abdel Magid et al. propose a concept-based auditing framework (Concept2Concept) that characterizes conditional prompt distributions through interpretable concepts, enabling systematic inspection of what a model tends to associate with a given prompt family and how those associations vary between prompt sets or usage scenarios [117]. This kind of concept-level view is particularly valuable when the goal is not only to produce visually plausible outputs but to also understand and control the semantic content that is implicitly “pulled in” by prompts during generation [117].

A complementary and more operational direction treats prompting as an iterative design process rather than a one-off instruction. Idea2Img formalizes this as an agent loop driven by a multi-modal model (GPT-4V): the system generates an initial prompt, synthesizes draught images, inspects the output, and then self-refines the prompt based on visually grounded feedback accumulated in memory about the behavior of the target T2I model [118]. This iterative procedure is motivated by the same practical reality that human

users already exploit, namely that the “effective prompt” is often discovered through interaction while making the refinement process explicit and automatable [118].

Beyond prompt engineering, several works highlight that T2I generation is increasingly conditioned by richer multi-modal signals and constraints. Wang et al. [116] study multi-modal style transfer guided by multi-modality, combining text-style and image-style guidance within a unified pipeline based on GAN inversion; in effect, the target “style” becomes a controllable multi-modal specification rather than a single reference image or a purely textual descriptor [116]. Although positioned in the style-transfer setting, the contribution is relevant to the practice of T2I because it illustrates how cross-modal conditioning can be used to steer the generation more precisely and to interpolate between multiple style anchors in a controlled way [116].

Finally, text-conditioned generation is increasingly used as an entry point to richer outputs that go beyond a single 2D image. Hong et al. introduce a zero-shot pipeline that uses natural language descriptions to guide not only appearance, but also geometry and motion, producing textured 3D avatars and plausible animations under CLIP-based guidance [119]. Although the outputs are 3D, the work is conceptually aligned with T2I developments in that it treats language as the primary interface for specifying visual content and extends the generation target toward more structured multi-attribute representations [119].

Taken together, these works suggest a consistent trend: modern “text-to-image” generation is best understood as a multi-modal system problem. It involves:

1. Managing underspecification and ambiguity as first-class properties [115];
2. Auditing and interpreting prompt-conditioned semantics [117];
3. Iterative, visually grounded prompt refinement loops [118];
4. Mechanisms of cross-modal conditioning for controllable generation [116];
5. The extension of text-conditioned generation to structured multi-attribute outputs such as 3D assets [119].

6.5. Image-to-Text Generation

Image-to-text generation constitutes a complementary direction to text-to-image modeling, focusing on transforming visual input into structured natural language outputs. Within the analyzed corpus, this line of work spans zero-shot captioning, cross-modal retrieval, assistive description systems, and context-aware or region-sensitive narrative generation, reflecting both methodological diversity and application-orientated motivations.

Early advances in zero-shot and prompt-based captioning are exemplified by [120], the authors of which introduced a training-free captioning framework that leverages pre-trained vision–language models. By guiding a frozen multi-modal encoder with textual prompts and optimization in the latent space, the method demonstrated that competitive captions can be generated without task-specific fine-tuning. This work established a baseline paradigm in which the cross-modal alignment learnt during large-scale pre-training can be directly exploited for image-to-text generation under minimal supervision.

Cross-modal retrieval and representation learning are further explored in [121], where semantic alignment between visual and textual modalities is treated as a bidirectional mapping problem. The study analyzes how shared embedding spaces facilitate not only retrieval but also descriptive synthesis, highlighting the importance of structured feature interaction and semantic consistency for reliable caption generation. By focusing on representation coherence, this line of research reinforces the foundational role of alignment quality in downstream generative tasks.

More application-driven perspectives emerge in inclusive and assistive description systems. In particular, Fernandes et al. [122] propose VIIDA, a computational framework to generate inclusive image paragraphs tailored to visually impaired users, alongside InViDe,

a dedicated evaluation metric grounded in accessibility criteria. By integrating multi-modal visual question answering with linguistic post-processing and scene graph structuring, the approach moves beyond short captions towards semantically organized paragraphs. This shift from brevity to structured inclusivity illustrates how image-to-text generation can be adapted to domain-specific communicative needs.

Region-sensitive and multi-scale reasoning is addressed in [100], the authors of which introduce mechanisms for progressively refining visual attention and descriptive granularity. Their approach demonstrates that hierarchical inspection of image regions enables more precise and context-aware captions, particularly in complex scenes where global summarization may obscure salient details. Such multi-level modeling strategies indicate a transition from flat captioning towards structured visual reasoning.

Taken together, these works reveal a gradual evolution of image-to-text generation from prompt-guided zero-shot captioning and embedding alignment techniques to more specialized, context-aware, and user-centric systems. Although foundational alignment remains central, recent contributions increasingly emphasize structural richness, inclusivity, and adaptive description strategies, thereby extending the scope of multi-modal generation beyond generic captioning toward functionally grounded communication.

6.6. Video and Temporal Multi-Modal Generation

Recent work on multi-modal generation has increasingly moved beyond static image–text settings to temporally structured audiovisual content, where motion, rhythm, and cross-modal synchronization become central design constraints. Two representative but conceptually distinct directions are illustrated by [123,124].

Luo et al. [123] present an art-driven multi-modal system in which dance video, ink-style visual transformation, and AI-generated music are integrated into a coherent montage. Their pipeline combines diffusion-based style transfer with LoRA-based fine-tuning to transform recorded dance sequences into a Chinese ink-painting aesthetic. In parallel, motion descriptors (e.g., pose key-points, joint velocities, and scene features) are extracted from video frames and used to condition a video-to-music generation module that predicts stylistically and rhythmically aligned musical structures. The contribution is not merely generative in isolation, but explicitly temporal and compositional: synchronization between movement dynamics and musical tempo is treated as a first-class objective, highlighting the importance of frame-level motion features in shaping audio generation. This work exemplifies how multi-modal diffusion and conditioning mechanisms can be extended from static alignment to culturally grounded, temporally evolving audiovisual artefacts.

In contrast, Wu et al. [124] focus on controllable temporal animation within a generative modeling framework. Their approach addresses the challenge of animating visual content under structured motion constraints, emphasizing consistency across frames and semantically meaningful motion trajectories. Rather than focusing on cross-artistic montage, the paper investigates how generative models can encode and propagate motion representations over time while preserving appearance coherence. The technical emphasis lies in disentangling motion and content factors and ensuring stable temporal dynamics during generation, thereby mitigating common issues such as flickering, identity drift, or inconsistent geometry.

Together, these works illustrate two complementary perspectives on video and multi-modal temporal generation. One direction emphasizes cross-modal synchronization and artistic composition in video and audio streams [123], while the other focuses on internal temporal coherence and motion control within generative visual pipelines [124]. Both underscore that extending multi-modal systems to the temporal domain requires explicit modeling of dynamics—either as cross-modal rhythm alignment or as structured motion

propagation—marking a shift from static multi-modal alignment toward fully spatio-temporal generative architectures.

6.7. Bias, Behavior, and Emergent Properties in Multi-Modal Models

Beyond architectural and generative advances, recent work has increasingly examined how multi-modal models internalize, amplify, or subtly reshape socially embedded biases. In this context, Aman et al. [126] provide one of the first systematic investigations of animal stereotyping in vision language models (VLMs), focusing on text-to-image generation. Through controlled prompting of DALL-E 3 with trait-animal templates (e.g., “wise animal”, “unfaithful animal”), the authors demonstrate highly concentrated associations between specific attributes and culturally established species (e.g., owls for wisdom, foxes for unfaithfulness). Importantly, the bias manifests not only at the level of species selection but also in visual depiction (facial expressions, posture, contextual background), indicating that stereotype reinforcement operates jointly across semantic and visual channels. The study further explores lightweight mitigation via prompt modification (“Do not stereotype animals”), showing partial diversification of outputs, yet confirming that such interventions do not fully eliminate bias rooted in training distributions.

Complementing this analysis from a broader multi-modal reasoning perspective, Chen et al. [125] examine whether multi-modal models reproduce or resist stereotypical associations when faced with socially sensitive demands. Their evaluation framework investigates behavioral tendencies under structured input variations, revealing that multi-modal systems can exhibit emergent patterns of alignment or misalignment depending on the task framework, the emphasis on the modality and contextual cues. Rather than being static artefacts, these behaviors appear to arise from complex interactions between pre-trained representations and prompt semantics, underscoring the dynamic and context-dependent nature of bias expression.

Taken together, these studies highlight that bias in multi-modal systems is not limited to textual prejudice inherited from large language models, nor solely to visual imbalance in generative diffusion frameworks. Instead, it emerges at the interface of modalities, where linguistic descriptors activate culturally encoded priors that are subsequently materialized through visual synthesis or multi-modal reasoning. This dual-layered phenomenon complicates both evaluation and mitigation, as interventions must address representational alignment, generative decoding, and prompt sensitivity simultaneously. Consequently, bias analysis in multi-modal models should be treated as a first-class research objective, tightly coupled with architectural design, dataset curation, and alignment strategies.

6.8. Multi-Lingual and Cross-Lingual Transfer in Multi-Modal Systems

Multi-lingual and cross-lingual multi-modal modeling extends generative frameworks beyond English-centric settings. It supports content generation, retrieval, and reasoning across diverse linguistic contexts. This cluster addresses both the scarcity of multi-lingual multi-modal data and the architectural adaptations needed for cross-lingual transfer.

An early and influential direction is represented by the generative cross-modal pre-training strategy proposed by Long et al. [127]. The authors introduce a unified framework in which visual and textual modalities are jointly modeled through generative objectives, allowing the model to learn language-agnostic semantic representations. By incorporating multi-lingual textual corpora into the pre-training pipeline, the approach facilitates cross-lingual transfer without requiring fully parallel multi-modal annotations for each target language. This generative perspective highlights how cross-modal pre-training can implicitly support multi-lingual generalization when aligned latent representations are sufficiently structured.

Complementing this line of work, Mohammed et al. [128] explicitly tackle the problem of limited multi-lingual multi-modal supervision. Their bootstrapping strategy leverages high-resource language–vision pairs to iteratively improve performance in lower-resource languages. By combining cross-lingual language modeling with visual grounding signals, the proposed framework demonstrates that multi-lingual transfer can be strengthened through shared multi-modal embeddings and weakly supervised data expansion. The study underscores the importance of representation sharing and iterative pseudo-labeling mechanisms in extending multi-modal systems beyond dominant languages.

Onami et al. [129], who investigate the answer to multi-lingual document questions in visually rich contexts, offer a more application-orientated perspective. Their approach integrates cross-lingual language models with document-level visual encoders, enabling question answering across multiple languages in structured document images. Unlike purely text-based cross-lingual QA systems, this framework must simultaneously handle visual layout, textual content, and language transfer. The results demonstrate that effective multi-lingual multi-modal reasoning requires coordinated alignment across textual, visual, and structural representations, particularly in document-centric scientific and technical domains.

Taken together, these studies illustrate three complementary mechanisms for multi-lingual and cross-lingual multi-modal transfer: (i) generative cross-modal pre-training with multi-lingual corpora [127], (ii) bootstrapped alignment and representation sharing across languages [128], and (iii) task-specific integration of cross-lingual language models with visual reasoning modules [129]. Collectively, they demonstrate that multi-lingual multi-modal systems benefit from unified latent spaces, data-efficient transfer strategies, and architectures explicitly designed to preserve semantic consistency across languages and modalities.

6.9. Section-Specific Limitations

The multi-modal and cross-lingual evidence base includes heterogeneous tasks (generation, alignment, captioning, retrieval, and document understanding) and uneven benchmark coverage across languages, which constrains strict cross-study comparability. Several works report strong results on specialized or emerging resources that are not yet broadly standardized, and some evaluations prioritize alignment quality without fully harmonized downstream utility criteria. As a result, conclusions in this section are best interpreted as a synthesis of robust thematic trajectories and recurrent bottlenecks, not as a definitive ranking across all multi-modal paradigms.

From a capability-centric perspective, the retained multi-modal corpus repeatedly exposes three coupled bottlenecks that are not resolved by raw model scale alone. First, *alignment strength versus grounding reliability*: systems optimized for high cross-modal similarity often improve benchmark retrieval and caption scores but remain brittle under region-level grounding or document-layout constraints, where local evidence tracking is required rather than global embedding agreement. Second, *unified architecture breadth versus controllable specialization*: any-to-any or foundation-style multi-modal models improve task coverage across modalities, but task-specific pipelines still outperform them in constrained settings that require explicit structural priors (e.g., document QA layouts, accessibility-oriented paragraph generation, or region-focused generation). Third, *English-centric pretraining efficiency versus multi-lingual robustness*: transfer works well for high-resource languages, yet performance stability degrades for lower-resource or domain-specific language distributions unless additional bootstrapping, curation, or language-aware alignment stages are introduced. These trade-offs indicate that next-generation multi-modal systems for scientific content workflows will likely need hybrid designs that combine shared latent

spaces with explicit grounding modules and language-adaptive calibration rather than relying exclusively on monolithic scaling.

To complement the section-specific syntheses, Table 5 summarizes representative datasets, benchmarks, and evaluation resources that recur across the reviewed literature and shape reproducibility, comparability, and downstream system design.

Table 5. Representative datasets, benchmarks, and evaluation resources referenced across the reviewed literature.

Reference	Year	Resource	Type/Scope	Review Relevance
<i>Generative Training and Domain Resources</i>				
Wang et al. [20]	2024	Text-of-gaze dataset	Dense textual gaze and head-pose descriptions for face control	Enables structured supervision for controllable face synthesis.
Li et al. [24]	2023	Matterport3D-conditioned VLN panoramas	Room descriptions paired with 360-degree navigation scenes	Grounds panoramic generation in realistic spatial settings.
Peng et al. [95]	2024	Anime3D	Anime-style single-image-to-3D benchmark data	Supports evaluation of 3D character reconstruction fidelity.
<i>Quality and Preference Benchmarks</i>				
Li et al. [60]	2023	AGIQA-3K	Human-annotated multi-dimensional benchmark dataset with MOS supervision and structured evaluation protocol.	Establishes multi-dimensional AIGIQA training and comparison.
Li et al. [61]	2024	AIGIQA-20K	Large-scale dataset expansion incorporating artifact-level annotations and structured training splits.	Expands large-scale supervision to improve robustness and generalization in AIGIQA modeling.
Yuan et al. [62]	2024	PKU-AIGIQA-4K	Unified perceptual dataset spanning text-to-image and image-to-image generative settings.	Extends perceptual benchmarking to multiple generative settings for broader QA evaluation.
Liu et al. [63]	2024	NTIRE 2024 AIGC Quality Assessment benchmarking framework with standardized tracks and evaluation criteria.	Establishes benchmarking standards and comparative baselines for AIGC quality assessment.	
Kirstain et al. [59]	2023	Pick-a-Pic/PickScore	Pairwise human preference benchmark	Anchors ranking evaluation to real-user preferences.
Wu et al. [57]	2023	HPS v2	Human preference scoring benchmark	Supports human-aligned model comparison.
<i>Multi-modal and Cross-Lingual Evaluation Resources</i>				
Chen et al. [111]	2024	Panda-70M	Large-scale paired multi-modal corpus	Supports robust vision–language alignment pre-training.
Fernandes et al. [122]	2025	VIIDA/InViDe	Inclusive image-description benchmark and metric	Extends evaluation toward accessibility-oriented generation.
Onami et al. [129]	2024	JDocQA	Japanese document QA benchmark	Highlights multi-lingual, document-centric evaluation needs for practical scientific workflows.

7. Discussion

General conclusions. The reviewed literature indicates that AI-driven image generation has moved from a collection of narrowly scoped modeling strategies into a broader ecosystem of multi-modal generative systems, evaluation frameworks, and domain-specific workflows. Across the architectural literature, the dominant trajectory is clear: GAN-based pipelines established the practical feasibility of high-quality synthesis, diffusion models improved fidelity and controllability, and transformer-based foundation models expanded generation into more unified text–image–video–3D settings. This progression is mirrored in the evaluation literature, where early distribution-level metrics have gradually been complemented by prompt-aware, perceptual, and human-centred assessment strategies better aligned with the actual use of generative systems.

Cross-cutting challenges. Controllability has become as important as raw perceptual quality. In the architectural sections, this appears as identity preservation, panoramic consistency, temporal coherence, and 3D structural stability. In the quality assessment

literature, it appears as the need to separately model realism, semantic faithfulness, authenticity, and user preference rather than collapsing them into a single scalar notion of quality. In the application-oriented literature, it appears as the practical requirement that systems remain editable, explainable, and responsive to constraints specific to design, storytelling, cultural heritage, and scientific communication. The resource landscape summarized in Table 5 makes a second challenge visible: reusable benchmarks and shared data resources are comparatively mature for AIGI quality assessment, but much less consolidated for application-centric, 3D, and multi-lingual multi-modal settings.

Capability-level synthesis. A comparison across the four review branches shows that the field is organized less by isolated model families than by recurring capability trade-offs. In generative architectures, improvements in fidelity and scale repeatedly introduce new constraints on controllability, inference cost, and compositional reliability. In evaluation research, benchmark performance remains useful only when it is interpreted alongside human preference, prompt faithfulness, annotation stability, and vulnerability to reward or metric overfitting. In application studies, practical value depends not only on image quality but also on whether systems expose enough structure for users to correct, constrain, and reuse generated outputs in domain-specific workflows. In multi-modal and cross-lingual settings, broader model unification improves coverage but does not automatically solve grounding, region-level reasoning, or language-specific robustness. Taken together, these patterns indicate that future progress will depend on hybrid designs that combine scalable generative backbones with explicit control mechanisms, transparent evaluation protocols, and deployment-oriented validation rather than on model scale alone.

Technology-specific bottlenecks. The main unresolved bottlenecks are not uniform across the field. Generative architecture research still struggles with compositional reasoning, identity preservation, temporal stability, and 3D consistency. Evaluation research is constrained by fragmented prompt sets, heterogeneous reporting conventions, and imperfect alignment between automatic metrics and human judgement. Application studies remain vulnerable to small user samples, narrow deployment contexts, and limited longitudinal validation, while multi-modal and cross-lingual systems continue to suffer from scarce benchmarks and under-specified bias evaluation protocols.

Limitations of the included evidence. The evidence base synthesized in this review is heterogeneous in publication type, evaluation design, and reporting depth. Many studies emphasize proof-of-concept performance, benchmark-specific gains, or narrowly scoped user studies rather than long-term comparative validation across shared protocols. The application literature is especially uneven with respect to sample sizes, deployment realism, and domain transferability, while multi-modal and cross-lingual studies often rely on emerging benchmarks whose coverage remains incomplete. These characteristics limit direct comparability across studies and make strong cross-sectional claims about maturity or superiority difficult.

Limitations of the review process. The review was based on a single primary bibliographic source (Scopus) and on a structured but non-meta-analytic narrative synthesis workflow. Thematic assignment and section-level refinement were designed to improve transparency and workload balance, but they also introduced rule-based and judgement-based decisions that may have affected borderline cases. In addition, the review was not prospectively registered, no separate archived protocol was prepared, and no formal study-level risk-of-bias or certainty-of-evidence framework was applied across the full corpus. The present synthesis does not include a bibliometric or visual-analytical overlay such as keyword co-occurrence mapping or co-citation network analysis; the structured four-branch thematic partitioning was judged to provide a sufficiently organized literature map for a structured narrative review of this scope, and a separate bibliometric layer would not materially

extend the narrative synthesis within the chosen reporting format. These choices do not prevent the review from offering a structured synthesis of the field, but they should be taken into account when interpreting the scope and reproducibility of the reported conclusions.

Future directions. Future research should therefore prioritize standardized prompt-aware benchmarks, stronger links between objective and human evaluation, and more realistic studies of how generative systems are used in professional and scientific workflows over time. Equally important will be the development of open, reusable resources beyond the quality assessment subfield, including multi-modal and multi-lingual benchmark suites, domain-specific application datasets, and evaluation protocols that jointly capture technical performance, usability, and trustworthiness.

8. Conclusions

This structured narrative review synthesized recent research on AI-driven image generation across four complementary dimensions: generative architectures, quality assessment and performance metrics, application domains, and multi-modal or cross-lingual extensions. The surveyed evidence shows that the field is no longer defined solely by the problem of generating visually plausible images. Instead, it is increasingly shaped by questions of controllability, semantic grounding, multi-modal coordination, human-centred evaluation, and practical deployment in complex real-world settings.

The strongest overall trend is the convergence of previously separate research streams. Modern generative pipelines increasingly combine diffusion, transformers, and foundation-model principles; evaluation is moving toward hybrid objective–subjective paradigms; and applications increasingly depend on systems that support iterative prompting, structural constraints, and domain-aware interaction. At the same time, the review highlights unresolved challenges related to benchmark comparability, compositional reasoning, identity and temporal consistency, multi-lingual robustness, and the gap between laboratory evaluation and applied use.

In summary, AI-driven image generation has entered a phase of methodological consolidation and practical expansion. The next stage of progress will likely depend on linking high-quality generation with robust evaluation, interpretable multi-modal reasoning, and trustworthy integration into domain-specific workflows. This makes future work on standardized benchmarks, prompt-aware assessment, multi-modal transfer, and user-centred system design especially important for the maturation of the field.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/electronics15132891/s1>. Table S1: PRISMA 2020 item-by-item reporting checklist (25 Yes, 0 Partial, 0 No, 12 N/A) with section and line references to the manuscript. Table S2: Review search, thematic assignment, branch-level screening, and retained-corpus summary, including record counts and artifact availability statement.

Author Contributions: Conceptualization, M.L.; methodology, M.L.; investigation, M.L., Y.Z., M.C.Q.F., D.M.C. and R.K.; writing—original draft preparation, M.L., Y.Z., M.C.Q.F., D.M.C. and R.K.; writing—review and editing, M.L., Y.Z., M.C.Q.F., D.M.C. and R.K.; supervision, M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by the program “Excellence Initiative—Research University” for the AGH University of Krakow. Research work was co-funded by the National Science Center, Poland, under decision DEC-2025/07/Y/ST6/00133 (MUTASK), as part of the CHIST-ERA project CHIST-ERA-25-SOL-06. This work was also supported in part by the National Natural Science Foundation of China under Grant 61901355, Grant 62271384, and Grant 62571418, and by the Japan Society for the Promotion of Science under Grant No. 22K12085.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The main study-characteristic and narrative-synthesis information is contained within the article and its Supplementary Review Documentation. The Supplementary Materials accompanying the manuscript include a PRISMA-oriented checklist mapping and a summary of review artifacts and workflow counts. Additional working materials, such as raw database export files and branch-level screening tables, are available from the corresponding author on reasonable request, subject to source-database licensing and submission constraints.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT 5 for the purposes of table generation, editing, etc. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

AIGI	AI-Generated Image
AIGQA	AI-Generated Image Quality Assessment
CLIP	Contrastive Language-Image Pre-training
DiT	Diffusion Transformer
FID	Fréchet Inception Distance
GAN	Generative Adversarial Network
LLM	Large Language Model
LoRA	Low-Rank Adaptation
MOS	Mean Opinion Score
NeRF	Neural Radiance Field
NR-IQA	No-Reference Image Quality Assessment
PLCC	Pearson Linear Correlation Coefficient
SRCC	Spearman Rank Correlation Coefficient
T2I	Text-to-Image
T2V	Text-to-Video
VLM	Vision–Language Model

References

1. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022.
2. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv* **2022**, arXiv:2204.06125.
3. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.L.; Ghasemipour, K.; Gontijo, Lopes, R.; Karagol, Ayan, B.; Salimans, T.; et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2022**, *35*, 36479–36494.
4. Zhang, L.; Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv* **2023**, arXiv:2302.05543.
5. Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; Aberman, K. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023.
6. Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; Rombach, R. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv* **2023**, arXiv:2307.01952.
7. Ye, H.; Zhang, J.; Liu, S.; Han, X.; Yang, W. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *arXiv* **2023**, arXiv:2308.06721.
8. Poole, B.; Jain, A.; Barron, J.T.; Mildenhall, B. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv* **2022**, arXiv:2209.14988.
9. Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; et al. CogView: Mastering Text-to-Image Generation via Transformers. In *Proceedings of the Advances in Neural Information Processing Systems*; Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34; pp. 19822–19835.

10. Team, C. Chameleon: Mixed-Modal Early-Fusion Foundation Models. *arXiv* **2024**, arXiv:2405.09818. <https://doi.org/10.48550/arXiv.2405.09818>.
11. Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In Proceedings of the 41st International Conference on Machine Learning (ICML), Vienna, Austria, 21–27 July 2024.
12. Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; et al. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv* **2025**, arXiv:2506.15742. <https://doi.org/10.48550/arXiv.2506.15742>.
13. Lian, L.; Li, B.; Yala, A.; Darrell, T. LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models. In Proceedings of the 12th International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
14. Kou, Z.; Pei, S.; Zhang, X. LeMon: Automating Portrait Generation for Zero-Shot Story Visualization with Multi-Character Interactions. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 25–29 August 2024. <https://doi.org/10.1145/3637528.3671850>.
15. Sheng, Y.; Tao, M.; Wang, J.; Bao, B.K. ISF-GAN: Imagine, Select, and Fuse with GPT-Based Text Enrichment for Text-to-Image Synthesis. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 1–7. <https://doi.org/10.1145/3650033>.
16. Li, W.; Li, H.; Peng, Y.; Wu, S.; Zhang, Y.; Sun, X. Create Anything Anywhere: Layout-Controllable Personalized Diffusion Model for Multiple Subjects. In Proceedings of the IEEE International Conference on Multimedia and Expo; IEEE: New York, NY, USA, 2025. <https://doi.org/10.1109/ICME59968.2025.11209272>.
17. Liu, Y.; An, J.; Chen, H.; Zhang, W.; Li, M.; Wu, D.; Gu, J.; Lin, Z.; Wang, W. Corer: Concept Residue Erasing in Text-to-Image Diffusion Models. In Proceedings of the IEEE International Conference on Multimedia and Expo; IEEE: New York, NY, USA, 2025. <https://doi.org/10.1109/ICME59968.2025.11210155>.
18. Praveen, D.; Harshitha, G.; Borkar, P.; D'souza, R.; Shetty, K.; Murthy, A. Innovative AI Solutions for Multi-Modal Content Generation. In Proceedings of the 2025 International Conference on Artificial Intelligence and Data Engineering (AIDE); IEEE: New York, NY, USA, 2025. <https://doi.org/10.1109/AIDE64228.2025.10987412>.
19. Kumar, J.; Tanvi, T.; Ghosh, A.; Mondal, R.; Das, S. Generating Images from Text Using GANs vs Transformers. In Proceedings of the 2025 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech); IEEE: New York, NY, USA, 2025. <https://doi.org/10.1109/IEMENTech65115.2025.10959693>.
20. Wang, H.; Zhang, Z.; Cheng, Y.; Chang, H. TextGaze: Gaze-Controllable Face Generation with Natural Language. In Proceedings of the 32nd ACM International Conference on Multimedia; ACM: New York, NY, USA, 2024. <https://doi.org/10.1145/3664647.3681252>.
21. Igmoullan, I.; El Boushaki, A.; El Bahi, H.; Hannane, R. Human Face Generation from Text Description and Sketch Using GANs and Transformers. In Proceedings of the International Conference on Computer Systems and Applications; IEEE: New York, NY, USA, 2025. <https://doi.org/10.1109/ICCSC66714.2025.11135126>.
22. Wang, Q.; Bai, X.; Wang, H.; Qin, Z.; Chen, A.; Li, H.; Tang, X.; Hu, Y. InstantID: Zero-shot Identity-Preserving Generation in Seconds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: New York, NY, USA, 2024.
23. Ye, W.; Ji, C.; Chen, Z.; Gao, J.; Huang, X.; Zhang, S.H.; Ouyang, W.; He, T.; Zhao, C.; Zhang, G. DiffPano: Scalable and Consistent Text to Panorama Generation with Spherical Epipolar-Aware Diffusion. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 1304–1332.
24. Li, J.; Bansal, M. PANOGEN: Text-Conditioned Panoramic Environment Generation for Vision-and-Language Navigation. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 21878–21894.
25. Zhang, C.; Wu, Q.; Gambardella, C.C.; Huang, X.; Phung, D.; Ouyang, W.; Cai, J. Taming Stable Diffusion for Text to 360 Panorama Image Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: New York, NY, USA, 2024; pp. 6347–6357.
26. Hong, W.; Ding, M.; Zheng, W.; Liu, X.; Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv* **2022**, arXiv:2205.15868.
27. Villegas, R.; Moraldo, H.; Castro, S.; Babaeizadeh, M.; Zhang, H.; Kunze, J.; Kindermans, P.J.; Saffar, M.; Erhan, D. Phenaki: Variable Length Video Generation from Open Domain Textual Descriptions. *arXiv* **2022**, arXiv:2210.02399.
28. Ma, X.; Wang, Y.; Chen, X.; Jia, G.; Liu, Z.; Li, Y.F.; Chen, C.; Qiao, Y. Latte: Latent Diffusion Transformer for Video Generation. *Trans. Mach. Learn. Res.* **2025**. Available online: <https://openreview.net/forum?id=vvnBP8u8v6> (accessed on 26 June 2026).
29. Menapace, W.; Siarohin, A.; Skorokhodov, I.; Deyneka, E.; Chen, T.S.; Kag, A.; Fang, Y.; Stolar, A.; Ricci, E.; Ren, J.; et al. Snap Video: Scaled Spatiotemporal Transformers for Text-to-Video Synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: New York, NY, USA, 2024.
30. Bar-Tal, O.; Chefer, H.; Tov, O.; Herrmann, C.; Paiss, R.; Zada, S.; Ephrat, A.; Hur, J.; Liu, G.; Raj, A.; et al. Lumiere: A Space-Time Diffusion Model for Video Generation. In SIGGRAPH Asia 2024 Conference Papers; ACM: New York, NY, USA, 2024.

31. Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; et al. Video generation models as world simulators. *OpenAI Blog* **2024**. Available online: <https://openai.com/research/video-generation-models-as-world-simulators> (accessed on 26 June 2026).
32. Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In Proceedings of the 13th International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025.
33. Kondratyuk, D.; Yu, L.; Gu, X.; Lezama, J.; Huang, J.; Schindler, G.; Hornung, R.; Birodkar, V.; Yan, J.; Chiu, M.C.; et al. VideoPoet: A Large Language Model for Zero-Shot Video Generation. *arXiv* **2023**, arXiv:2312.14125.
34. Li, P.; Chen, K.; Liu, Z.; Gao, R.; Hong, L.; Yeung, D.Y.; Lu, H.; Jia, X. TrackDiffusion: Tracklet-Conditioned Video Generation via Diffusion Models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2025. <https://doi.org/10.1109/WACV61041.2025.00349>.
35. Su, S.; Liu, J.; Gao, L.; Song, J. F3-Pruning: A Training-Free and Generalized Pruning Strategy towards Faster and Finer Text-to-Video Synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI: Washington, DC, USA, 2024. <https://doi.org/10.1609/aaai.v38i5.28300>.
36. Assran, M.; Bardes, A.; Fan, D.; Garrido, Q.; Howes, R.; Komeili, M.; Muckley, M.; Rizvi, A.; Roberts, C.; Sinha, K.; et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. *arXiv* **2025**, arXiv:2506.09985.
37. Zhang, S.; Zhang, Y.; Zheng, Q.; Ma, R.; Hua, W.; Bao, H.; Xu, W.; Zou, C. 3D-SceneDreamer: Text-Driven 3D-Consistent Scene Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024. <https://doi.org/10.1109/CVPR52733.2024.00969>.
38. Feng, W.; Zhu, W.; Fu, T.J.; Jampani, V.; Akula, A.; He, X.; Basu, S.; Wang, X.; Wang, W. LayoutGPT: Compositional Visual Planning and Generation with Large Language Models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 18225–18250.
39. Sargent, K.; Koh, J.; Zhang, H.; Chang, H.; Herrmann, C.; Srinivasan, P.; Wu, J.; Sun, D. VQ3D: Learning a 3D-Aware Generative Model on ImageNet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2023. <https://doi.org/10.1109/ICCV51070.2023.00391>.
40. Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; Tan, H. LRM: Large Reconstruction Model for Single Image to 3D. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
41. He, Y.; Zhou, Y.; Zhao, W.; Wu, Z.; Xiao, K.; Yang, W.; Liu, Y.J.; Han, X. StdGEN: Semantic-Decomposed 3D Character Generation from Single Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2025. <https://doi.org/10.1109/CVPR52734.2025.02453>.
42. Shi, Y.; Wang, P.; Ye, J.; Mai, L.; Li, K.; Yang, X. MVDream: Multi-view Diffusion for 3D Generation. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
43. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
44. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
45. Zhou, T.; Tan, S.; Li, L.; Zhao, B.; Jiang, Q.; Yue, G. Cross-Modality Interactive Attention Network for AI-Generated Image Quality Assessment. *Pattern Recognit.* **2025**, *167*, 111693. <https://doi.org/10.1016/j.patcog.2025.111693>.
46. Zhou, T.; Tan, S.; Zhou, W.; Luo, Y.; Wang, Y.G.; Yue, G. Adaptive Mixed-Scale Feature Fusion Network for Blind AI-Generated Image Quality Assessment. *IEEE Trans. Broadcast.* **2024**, *70*, 833–843. <https://doi.org/10.1109/TBC.2024.3391060>.
47. Yuan, J.; Cao, X.; Che, J.; Wang, Q.; Liang, S.; Ren, W.; Lin, J.; Cao, X. TIER: text-image Encoder-based Regression for AIGC Image Quality Assessment. *arXiv* **2024**, arXiv:2401.03854. <https://doi.org/10.48550/arXiv.2401.03854>.
48. Zhang, Y.; Jia, M. Quality Assessment of AI-Generated Image Based on Cross-Modal Correlation. In *Proceedings of the 3rd International Conference on Image Processing and Media Computing (ICIPMC)*; IEEE: New York, NY, USA, 2024. <https://doi.org/10.1109/ICIPMC62364.2024.10586696>.
49. Yu, Z.; Guan, F.; Lu, Y.; Li, X.; Chen, Z. SF-IQA: Quality and Similarity Integration for AI Generated Image Quality Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*; IEEE: New York, NY, USA, 2024; pp. 6692–6701.
50. Tang, Z.; Wang, Z.; Peng, B.; Dong, J. CLIP-AGIQA: Boosting the Performance of AI-Generated Image Quality Assessment with CLIP. In *Proceedings of the Pattern Recognition: 27th International Conference, ICPR 2024, Kolkata, India, 1–5 December 2024; Part XXXII*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 48–61. https://doi.org/10.1007/978-3-031-78125-4_4.
51. Chen, J.; Shao, F.; Chen, H.; Wang, X.; Guo, H.; Jiang, Q. Quality Evaluation of AI-Generated Images: Subjective Study and Objective Methodology. *IEEE Trans. Multimed.* **2026**, *28*, 57–70. <https://doi.org/10.1109/TMM.2025.3618562>.
52. Aziz, M.; Rehman, U.; Danish, M.U.; Ahmed, S.; Khan, M.; Ali, R. Towards a Unified Evaluation Framework: Integrating Human Perception and Metrics for AI-Generated Images. *Multimed. Syst.* **2025**, *31*, 112. <https://doi.org/10.1007/s00530-025-01769-7>.

53. Sridhar, Y.; Gajendra, J.M.; Srinivasalu, P.; Srinivasappa, M. Evaluating Non-Reference Image Quality Metrics for AI-Generated Images: A Novel Approach. *Eur. J. Artif. Intell.* **2025**, *4*, 18–25. <https://doi.org/10.24018/ejai.2025.4.5.70>.
54. Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; Choi, Y. CLIPScore: A Reference-Free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 7514–7528.
55. Hu, Y.; Liu, B.; Kasai, J.; Wang, Y.; Ostendorf, M.; Krishna, R.; Smith, N.A. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023. <https://doi.org/10.1109/ICCV51070.2023.01866>.
56. Ghosh, D.; Hajishirzi, H.; Schmidt, L. GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment. In *Proceedings of the Advances in Neural Information Processing Systems*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 52132–52152. <https://doi.org/10.52202/075280-2270>.
57. Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; Li, H. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *arXiv* **2023**, arXiv:2306.09341. <https://doi.org/10.48550/arXiv.2306.09341>.
58. Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; Dong, Y. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In *Proceedings of the Advances in Neural Information Processing Systems*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 15903–15935. <https://doi.org/10.52202/075280-0700>.
59. Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; Levy, O. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. In *Proceedings of the Advances in Neural Information Processing Systems*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 36652–36663. <https://doi.org/10.52202/075280-1594>.
60. Li, C.; Zhang, Z.; Wu, H.; Sun, W.; Min, X.; Liu, X.; Zhai, G.; Lin, W. AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 5483–5497. <https://doi.org/10.1109/TCSVT.2023.3319020>.
61. Li, C.; Kou, T.; Gao, Y.; Cao, Y.; Sun, W.; Zhang, Z.; Zhou, Y.; Zhang, Z.; Zhang, W.; Wu, H.; et al. AIGIQA-20K: A Large Database for AI-Generated Image Quality Assessment. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: New York, NY, USA, 2024; pp. 6327–6336. <https://doi.org/10.1109/CVPRW63382.2024.00636>.
62. Yuan, J.; Yang, F.; Li, J.; Cao, X.; Che, J.; Lin, J.; Cao, X. PKU-AIGIQA-4K: A Perceptual Quality Assessment Database for Both Text-to-Image and Image-to-Image AI-Generated Images. In *Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*; IEEE: New York, NY, USA, 2025; pp. 3362–3371. <https://doi.org/10.1109/ICCVW69036.2025.00353>.
63. Liu, X.; Min, X.; Zhai, G.; Li, C.; Kou, T.; Sun, W.; Wu, H.; Gao, Y.; Cao, Y.; Zhang, Z.; et al. NTIRE 2024 Quality Assessment of AI-Generated Content Challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*; IEEE: New York, NY, USA, 2024; pp. 6337–6362. <https://doi.org/10.1109/CVPRW63382.2024.00637>.
64. Hartwig, S.; Engel, D.; Sick, L.; Kniesel, H.; Payer, T.; Poonam, P.; Glöckler, M.; Bäuerle, A.; Ropinski, T. A Survey on Quality Metrics for Text-to-Image Generation. *IEEE Trans. Vis. Comput. Graph.* **2025**, *31*, 9464–9483. <https://doi.org/10.1109/TVCG.2025.3585077>.
65. Ghildyal, A.; Chen, Y.; Zadtootaghaj, S.; Barman, N.; Bovik, A.C. Quality Prediction of AI Generated Images and Videos: Emerging Trends and Opportunities. *arXiv* **2024**, arXiv:2410.08534. <https://doi.org/10.48550/arXiv.2410.08534>.
66. Tian, Y.; Li, Y.; Chen, B.; Zhu, H.; Wang, S.; Kwong, S. AI-Generated Image Quality Assessment in Visual Communication. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-25)*; AAAI: Washington, DC, USA, 2025; Volume 39; pp. 7392–7400. <https://doi.org/10.1609/aaai.v39i7.32795>.
67. Vinothkumar, S.; Varadhaganapathy, S.; Shanthakumari, R.; Dhanushya, S.; Guhan, S.; Krisvanth, P. Utilizing Generative AI for Text-to-Image Generation. In *Proceedings of the 15th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*; IEEE: New York, NY, USA, 2024; pp. 1–6. <https://doi.org/10.1109/ICCCNT61001.2024.10725454>.
68. Chen, J.; An, J.; Lyu, H.; Kanan, C.; Luo, J. Learning to Evaluate the Artness of AI-Generated Images. *IEEE Trans. Multimed.* **2024**, *26*, 10731–10740. <https://doi.org/10.1109/TMM.2024.3410672>.
69. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. <https://doi.org/10.1109/TIP.2012.2214050>.
70. He, W.; Xiao, Y.; Xie, Y. Revealing user tacit knowledge: Generative-Image-AI helps create better design conversation. In *Proceedings of the DRS2024, Boston, MA, USA, 23–28 June 2024*; Design Research Society: London, UK, 2024. <https://doi.org/10.21606/drs.2024.1329>.
71. Choi, D.; Hong, S.; Park, J.; Chung, J.J.Y.; Kim, J. CreativeConnect: Supporting Reference Recombination for Graphic Design Ideation with Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2024; pp. 1–25. <https://doi.org/10.1145/3613904.3642794>.

72. Chen, L.; Jing, Q.; Tsang, Y.; Wang, Q.; Sun, L.; Luo, J. DesignFusion: Integrating Generative Models for Conceptual Design Enrichment. *J. Mech. Des.* **2024**, *146*, 111703. <https://doi.org/10.1115/1.4065487>.
73. Duan, R.; Zhu, C.; Chen, Y.; Hu, Y.; Shi, J.; Ramani, K. DesignFromX: Empowering Consumer-Driven Design Space Exploration through Feature Composition of Referenced Products. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*; ACM: New York, NY, USA, 2025; pp. 1040–1060. <https://doi.org/10.1145/3715336.3735824>.
74. Chen, C.; Nguyen, C.; Groueix, T.; Kim, V.G.; Weibel, N. MemoVis: A GenAI-Powered Tool for Creating Companion Reference Images for 3D Design Feedback. *ACM Trans. Comput.-Hum. Interact.* **2024**, *31*, 67. <https://doi.org/10.1145/3694681>.
75. Lu, Y.; Yuan, W.; Wang, M.; Wang, P.; Wu, S.; Wu, J.; Xing, W.; Xie, W.; Yu, F. Multi-agent collaborative pathways for Chinese traditional architectural image generation. *Sci. Rep.* **2025**, *15*, 34596. <https://doi.org/10.1038/s41598-025-18130-7>.
76. Guida, G. multi-modal Architecture: Applications of Language in a Machine Learning Aided Design Process. In *Proceedings of the 28th Conference on Computer Aided Architectural Design Research in Asia (CAADRIA)*; CAADRIA: Hong Kong, China, 2023; Volume 2, pp. 561–570. <https://doi.org/10.52842/conf.caadria.2023.2.561>.
77. Paananen, V.; Oppenlaender, J.; Visuri, A. Using text-to-image generation for architectural design ideation. *Int. J. Archit. Comput.* **2023**, *22*, 458–474. <https://doi.org/10.1177/14780771231222783>.
78. Guo, Y.; Sun, L. Garment pattern generation method based on Diffusion Model and 3D reconstruction technology. *Text. Res. J.* **2025**, *95*, 2866–2881. <https://doi.org/10.1177/00405175251328296>.
79. Rizzi, G.; Bertola, P. Exploring the generative AI potential in the fashion design process: an experimental experience on the collaboration between fashion design practitioners and generative AI tools. *Eur. J. Cult. Manag. Policy* **2025**, *15*, 13875. <https://doi.org/10.3389/ejcmp.2025.13875>.
80. Sun, Y.; Chen, Y.; Wang, Z.; Ye, Q.; Lyu, Y.; Liu, H. Fashion style transfer: Enhancing efficiency and aesthetic with a novel two-stage neural network. *J. Text. Inst.* **2024**, *116*, 2075–2086. <https://doi.org/10.1080/00405000.2024.2413727>.
81. Ahn, N.; Lee, J.; Lee, C.; Kim, K.; Kim, D.; Nam, S.H.; Hong, K. DreamStyler: Paint by Style Inversion with Text-to-Image Diffusion Models. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 674–681. <https://doi.org/10.1609/aaai.v38i2.27824>.
82. Azuaje, G.; Liew, K.; Epure, E.; Yada, S.; Wakamiya, S.; Aramaki, E. Visualyre: multi-modal album art generation for independent musicians. *Pers. Ubiquitous Comput.* **2023**, *27*, 1861–1872. <https://doi.org/10.1007/s00779-023-01724-1>.
83. Ko, H.K.; Park, G.; Jeon, H.; Jo, J.; Kim, J.; Seo, J. Large-scale Text-to-Image Generation Models for Visual Artists' Creative Works. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*; ACM: New York, NY, USA, 2023; pp. 919–933. <https://doi.org/10.1145/3581641.3584078>.
84. Almeda, S.G.; Zamfirescu-Pereira, J.; Kim, K.W.; Mani Rathnam, P.; Hartmann, B. Prompting for Discovery: Flexible Sense-Making for AI Art-Making with DreamSheets. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2024; pp. 1–17. <https://doi.org/10.1145/3613904.3642858>.
85. Hod, G.; Ben Or, D.; Regev, E. From a Single Design to Multiple Variations: AI-Guided in-App Image Generation. In *Proceedings of the 2025 IEEE Conference on Games (CoG)*; IEEE: New York, NY, USA, 2025; pp. 1–8. <https://doi.org/10.1109/cog64752.2025.1114405>.
86. Leininger, P.; Weber, C.J.; Rothe, S. Understanding Creative Potential and Use Cases of AI-Generated Environments for Virtual Film Productions: Insights from Industry Professionals. In *Proceedings of the 2025 ACM International Conference on Interactive Media Experiences*; ACM: New York, NY, USA, 2025; pp. 60–78. <https://doi.org/10.1145/3706370.3727853>.
87. Fernandes, T.; Nisi, V.; Nunes, N.; James, S., ArtAI4DS: AI Art and Its Empowering Role in Digital Storytelling. In *Entertainment Computing—ICEC 2024*; Springer: Cham, Switzerland, 2024; pp. 78–93. https://doi.org/10.1007/978-3-031-74353-5_6.
88. Shahriyar, A.; Hamed, R.; Perkoff, E.M.; Aboelnaga, M.; Azab, A. BIBLIOSMIA: Hyper-Personalized Consistent Stories for Enhanced Social Emotional Learning. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*; Wang, Z., Woodhead, S., Ananda, M., Mallick, D.B., Sharpnack, J., Burstein, J., Eds.; PMLR: New York, NY, USA, 2025; Volume 273, pp. 105–115.
89. Liang, Z.; Zhang, X.; Ma, K.; Liu, Z.; Ren, X.; Goucher-Lambert, K.; Liu, C. StoryDiffusion: How to Support UX Storyboarding with Generative-AI. In *Proceedings of the 27th International Conference on Multimodal Interaction*; ACM: New York, NY, USA, 2025; pp. 135–144. <https://doi.org/10.1145/3716553.3750793>.
90. Yi Chan, E.M.; Seow, C.K.; Wee Tan, E.S.; Wang, M.; Yau, P.C.; Cao, Q. SketchBoard: Sketch-Guided Storyboard Generation for Game Characters in the Game Industry. In *Proceedings of the 2024 IEEE 22nd International Conference on Industrial Informatics (INDIN)*; IEEE: New York, NY, USA, 2024; pp. 1–8. <https://doi.org/10.1109/indin58382.2024.10774282>.
91. Numan, N.; Rajaram, S.; Kumaravel, B.T.; Marquardt, N.; Wilson, A.D. SpaceBlender: Creating Context-Rich Collaborative Spaces Through Generative 3D Scene Blending. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*; ACM: New York, NY, USA, 2024; pp. 1–25. <https://doi.org/10.1145/3654777.3676361>.
92. Epstein, D.; Poole, B.; Mildenhall, B.; Efros, A.A.; Holynski, A. Disentangled 3D scene generation with layout learning. *arXiv* **2024**, arXiv:2402.16936.

93. Zhang, J.; Li, X.; Wan, Z.; Wang, C.; Liao, J. Text2NeRF: Text-Driven 3D Scene Generation with Neural Radiance Fields. *IEEE Trans. Vis. Comput. Graph.* **2024**, *30*, 7749–7762. <https://doi.org/10.1109/tvcg.2024.3361502>.
94. Pu, G.; Zhao, Y.; Lian, Z. Pano2Room: Novel View Synthesis from a Single Indoor Panorama. In *Proceedings of the SIGGRAPH Asia 2024 Conference Papers*; ACM: New York, NY, USA, 2024; pp. 1–11. <https://doi.org/10.1145/3680528.3687616>.
95. Peng, H.Y.; Zhang, J.P.; Guo, M.H.; Cao, Y.P.; Hu, S.M. CharacterGen: Efficient 3D Character Generation from Single Images with Multi-View Pose Canonicalization. *ACM Trans. Graph.* **2024**, *43*, 84. <https://doi.org/10.1145/3658217>.
96. Dong, J.; Fang, Q.; Huang, Z.; Xu, X.; Wang, J.; Peng, S.; Dai, B.; TELA: Text to Layer-Wise 3D Clothed Human Generation. In *Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024; pp. 19–36. https://doi.org/10.1007/978-3-031-72698-9_2.
97. Yamada, F.M.; Takahashi, H. AniDream: Generating Skeleton-Guided Anime Avatars from Text Prompts. In *Proceedings of the 2025 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*; IEEE: New York, NY, USA, 2025; pp. 1042–1052. <https://doi.org/10.1109/ismar67309.2025.00111>.
98. Li, D.Y.; Liu, Y.L.; Liu, Z.X.; Cao, Y.P.; Guo, M.H.; Hu, S.M. SeparateGen: Semantic Component-based 3D Character Generation from Single Images. *IEEE Trans. Vis. Comput. Graph.* **2026**, *32*, 4973–4986. <https://doi.org/10.1109/tvcg.2026.3652452>.
99. Lin, W.; Karlinsky, L.; Shvetsova, N.; Possegger, H.; Kozłowski, M.; Panda, R.; Feris, R.; Kuehne, H.; Bischof, H. MATCH, eXpand and Improve: Unsupervised Finetuning for Zero-Shot Action Recognition with Language Knowledge. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2023. <https://doi.org/10.1109/ICCV51070.2023.00267>.
100. Ramu, P.; Gaur, P.; Emandi, R.; Maheshwari, H.; Javed, D.; Garimella, A. Zooming in on Zero-Shot Intent-Guided and Grounded Document Generation using LLMs. In *Proceedings of the 17th International Natural Language Generation Conference, Tokyo, Japan*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 676–694. <https://doi.org/10.18653/v1/2024.inlg-main.52>.
101. Cho, J.; Lu, J.; Schwenk, D.; Hajishirzi, H.; Kembhavi, A. X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 8785–8805. <https://doi.org/10.18653/v1/2020.emnlp-main.707>.
102. Yu, L.; Cheng, Y.; Wang, Z.; Kumar, V.; Macherey, W.; Huang, Y.; Ross, D.; Essa, I.; Bisk, Y.; Yang, M.H.; et al. SPAE: Semantic Pyramid AutoEncoder for multi-modal Generation with Frozen LLMs. In *Proceedings of the Advances in Neural Information Processing Systems*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 52692–52704. <https://doi.org/10.52202/075280-2294>.
103. Fang, Y.; Khademi, M.; Zhu, C.; Yang, Z.; Pryzant, R.; Xu, Y.; Qian, Y.; Yoshioka, T.; Yuan, L.; Zeng, M.; et al. i-Code Studio: A Configurable and Composable Framework for Integrative AI. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Miami, FL, USA*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 14–24. <https://doi.org/10.18653/v1/2024.emnlp-demo.2>.
104. Jain, J.; Yang, J.; Shi, H. VCoder: Versatile Vision Encoders for multi-modal Large Language Models. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024. <https://doi.org/10.1109/CVPR52733.2024.02644>.
105. Pan, K.; Tang, S.; Li, J.; Fan, Z.; Chow, W.; Yan, S.; Chua, T.S.; Zhuang, Y.; Zhang, H. Auto-Encoding Morph-Tokens for multi-modal LLM. In *Proceedings of the 41st International Conference on Machine Learning*; Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F., Eds.; PMLR: New York, NY, USA, 2024; Volume 235, pp. 39308–39323.
106. Tang, Z.; Yang, Z.; Khademi, M.; Liu, Y.; Zhu, C.; Bansal, M. CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024. <https://doi.org/10.1109/CVPR52733.2024.02589>.
107. Huang, Y.; Qi, H.; Chen, Z.; Zhang, H.; Yu, H.; Zhao, Z. Autonomous multi-modal Reasoning via Implicit Chain-of-Vision. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*; IEEE: New York, NY, USA, 2025. <https://doi.org/10.1109/CVPRW67362.2025.00279>.
108. Cui, Y.; Sun, Q.; Zhang, X.; Zhang, F.; Yu, Q.; Luo, Z.; Wang, Y.; Rao, Y.; Liu, J.; Huang, T.; et al. Generative multi-modal Models Are In-Context Learners. In *Large Vision–Language Models. Advances in Computer Vision and Pattern Recognition*; Springer: Cham, Switzerland, 2026. https://doi.org/10.1007/978-3-031-94969-2_4.
109. Wu, H.H.; Seetharaman, P.; Kumar, K.; Bello, J. Way2CLIP: Learning Robust Audio Representations from Clip. In *Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*; IEEE: New York, NY, USA, 2022. <https://doi.org/10.1109/ICASSP43922.2022.9747669>.
110. Yang, S.; Li, Q.; Li, W.; Li, X.Y.; Jin, R.; Lv, B.; Wang, R.; Liu, A. Semantic Completion and Filtration for image–text Retrieval. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 140. <https://doi.org/10.1145/3572844>.
111. Chen, T.S.; Siarohin, A.; Menapace, W.; Deyneka, E.; Chao, H.W.; Jeon, B.; Fang, Y.; Lee, H.Y.; Ren, J.; Yang, M.H.; et al. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024. <https://doi.org/10.1109/CVPR52733.2024.01265>.

112. Fang, S.; Wu, Y.; Gao, B.; Cai, J.; Teik, T. Exploring Efficient-Tuned Learning Audio Representation Method from BriVL. In *Communications in Computer and Information Science*; Springer: Singapore, 2024. https://doi.org/10.1007/978-981-99-8184-7_4.
113. Bansal, H.; Bitton, Y.; Szepkator, I.; Chang, K.W.; Grover, A. VideoCon: Robust Video-Language Alignment via Contrast Captions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024. <https://doi.org/10.1109/CVPR52733.2024.01321>.
114. Campi, R.; Borrego, S.; De Santis, A.; Bianchi, M.; Tocchetti, A.; Brambilla, M. Towards Synthetic Concept Activation Vectors via Generative Models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*; IEEE: New York, NY, USA, 2025. <https://doi.org/10.1109/CVPRW67362.2025.00256>.
115. Hutchinson, B.; Baldridge, J.; Prabhakaran, V. Underspecification in Scene Description-to-Depiction Tasks. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 1172–1184. <https://doi.org/10.18653/v1/2022.aacl-main.86>.
116. Wang, H.; Wu, P.; Dela Rosa, K.; Wang, C.; Shrivastava, A. multi-modality-guided Image Style Transfer using Cross-modal GAN Inversion. In *Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2024; pp. 4964–4973. <https://doi.org/10.1109/WACV57701.2024.00490>.
117. Abdel Magid, S.; Pan, W.; Warchol, S.; Guo, G.; Kim, J.; Rahman, M.; Pfister, H. Is What You Ask for What You Get? Investigating Concept Associations in Text-to-Image Models. *arXiv* **2024**, arXiv:2410.04634.
118. Yang, Z.; Wang, J.; Li, L.; Lin, K.; Lin, C.C.; Liu, Z.; Wang, L. Idea2Img: Iterative Self-refinement with GPT-4V for Automatic Image Design and Generation. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2025. https://doi.org/10.1007/978-3-031-72920-1_10.
119. Hong, F.; Zhang, M.; Pan, L.; Cai, Z.; Yang, L.; Liu, Z. Text-Conditioned Zero-Shot 3D Avatar Creation and Animation. In *Advances in Computer Vision and Pattern Recognition*; Springer: Cham, Switzerland, 2026. https://doi.org/10.1007/978-3-031-94969-2_15.
120. Tewel, Y.; Shalev, Y.; Schwartz, I.; Wolf, L. ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022. <https://doi.org/10.1109/CVPR52688.2022.01739>.
121. Żelaszczyk, M.; Mańdziuk, J. Cross-modal text and visual generation: A systematic review. Part 1: Image to text. *Inf. Fusion* **2023**, *93*, 302–329. <https://doi.org/10.1016/j.inffus.2023.01.008>.
122. Fernandes, D.; Ribeiro, M.; Silva, M.; Cerqueira, F. VIIDA and InViDe: Computational approaches for generating and evaluating inclusive image paragraphs for the visually impaired. *Disabil. Rehabil. Assist. Technol.* **2025**, *20*, 1470–1495. <https://doi.org/10.1080/17483107.2024.2437567>.
123. Luo, M.; Zhang, Y.; Xu, P.; Wang, T.; Bo, Y.; Jin, X.; Dong, W. Dance Montage through Style Transfer and Music Generation. In *Proceedings of the SIGGRAPH Asia 2024 Art Papers*; ACM: New York, NY, USA, 2024; pp. 1–5. <https://doi.org/10.1145/3680530.3695444>.
124. Wu, X.; Huang, Z.; Yu, C. Animating the Past: Reconstruct Trilobite via Video Generation. In *Proceedings of the 2024 IEEE International Conference on Big Data (BigData)*; IEEE: New York, NY, USA, 2024; pp. 3333–3342. <https://doi.org/10.1109/BigData62323.2024.10825790>.
125. Chen, T.; Hirota, Y.; Otani, M.; García, N.; Nakashima, Y. Would Deep Generative Models Amplify Bias in Future Models? In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024. <https://doi.org/10.1109/CVPR52733.2024.01030>.
126. Aman, T.; Nadeem, M.; Sohail, S.; Anas, M.; Cambria, E. Owls are Wise and Foxes are Unfaithful: Uncovering Animal Stereotypes in Vision Language Models. In *Lecture Notes in Computer Science*; Springer: Singapore, 2025. https://doi.org/10.1007/978-981-96-8197-6_31.
127. Long, Q.; Wang, M.; Li, L. Generative Imagination Elevates Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 5738–5748. <https://doi.org/10.18653/v1/2021.naacl-main.457>.
128. Mohammed, O.K.; Aggarwal, K.; Liu, Q.; Singhal, S.; Bjorck, J.; Som, S. Bootstrapping a High Quality multi-lingual multi-modal Dataset for Bletchley. In *Proceedings of the 14th Asian Conference on Machine Learning*; Khan, E., Gonen, M., Eds.; PMLR: New York, NY, USA, 2023; Volume 189, pp. 738–753.
129. Onami, E.; Kurita, S.; Miyanishi, T.; Watanabe, T. JDocQA: Japanese Document Question Answering Dataset for Generative Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia; European Language Resources Association (ELRA): Paris, France; ICCL: Stroudsburg, PA, USA, 2024; pp. 9503–9514. <https://doi.org/10.63317/5k3radukw3sc>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.