

Review

Collaborative Control in Diffusion Models for Precise Image Generation: A Survey

Jingzhong Qi ¹, Wei Xu ¹, Qing Zhu ^{1,2}, Xinchen Chu ¹ and Yifan Wang ^{1,*}

¹ College of Computer Science, Beijing University of Technology, Chaoyang District, Beijing 100124, China; qijingzhong@emails.bjut.edu.cn (J.Q.); weixu@bjut.edu.cn (W.X.); ccgszq@bjut.edu.cn (Q.Z.); xinchenchu1234@emails.bjut.edu.cn (X.C.)

² Beijing Institute of Artificial Intelligence, Chaoyang District, Beijing 100124, China

* Correspondence: yifanewang@emails.bjut.edu.cn

Abstract

Diffusion models have become a central paradigm for image generation because they combine stable optimization, high-fidelity synthesis, and controllability through iterative denoising. However, precise image generation in practical settings requires more than text prompts. Here, precision means measurable satisfaction of semantic, spatial, structural, identity, interaction, and domain constraints rather than pixelwise reproduction alone. This survey examines collaborative control methods for diffusion-based image generation from a system-level perspective. We distinguish ordinary controllable diffusion from collaborative control, then review theoretical foundations, conditional generation, single-condition extensions, multi-condition fusion, conflict mediation, controller–evaluator loops, scalability, applications, and evaluation protocols. The discussion emphasizes how control signals are represented, injected, scheduled, and evaluated along the denoising trajectory. It also compares representative methods in terms of controllability, computational overhead, scalable inference, and task-oriented metrics. We identify three continuing challenges: robust coordination under conflicting heterogeneous conditions, fine-grained control under few-step sampling, and reliable benchmarks that jointly measure constraint satisfaction and efficiency. Overall, collaborative control reframes precise diffusion generation as a coordinated modeling and optimization problem involving models, conditions, schedulers, and evaluators.

Keywords: diffusion models; image generation; controllable generation; intelligent computing

MSC: 68T07; 68T45



Academic Editor: Aydin Azizi

Received: 24 June 2026

Revised: 20 July 2026

Accepted: 22 July 2026

Published: 2 August 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

The rapid development of large-scale visual data, multimodal learning, and generative artificial intelligence has made reliable and controllable image generation a central research problem. Diffusion models have become a leading solution because they combine stable optimization, expressive distribution modeling, and high-quality synthesis [1,2]. Their iterative denoising process also provides a natural framework for constraint-aware synthesis [3]. In practical scenarios, a text prompt is rarely sufficient. Target images may need to satisfy layout, geometry, object relations, identity, interaction, and domain rules supplied by masks, depth maps, poses, reference images, or external knowledge.

In this survey, *precise image generation* does not mean exact pixel reproduction in every task. It means that an output satisfies a declared set of measurable constraints at the granularity required by the application. These constraints may be pixel-level, such as edge

or mask agreement; structural, such as depth, pose, or layout; semantic, such as object count and attribute binding; identity-related; or task-specific. Precision is therefore operational and multidimensional. A system is precise only with respect to stated constraints, tolerances, and metrics, while perceptual quality and diversity remain separate objectives.

The literature has been organized through complementary perspectives, including the evolution of text-to-image backbones, conditioning signals, controllable diffusion mechanisms, and application domains [4,5]. Building on these perspectives, this survey focuses on the system-level question of how several conditions and control components interact within one denoising process. This focus requires condition conflicts, timestep scheduling, external planning, evaluator feedback, and computational cost to be analysed as connected parts of the same control problem.

The distinguishing feature of this survey is a control-loop organization of precise image generation. We connect six stages: condition representation, injection into the denoising backbone, multi-condition fusion and arbitration, timestep-dependent scheduling, evaluator-guided feedback, and deployment-oriented assessment. This organization links individual control mechanisms to the conflicts they address, the stage at which they act, and the computational or robustness trade-offs they introduce. It also provides one framework for comparing branch-based controllers, lightweight adapters, regional or multi-path inference, dynamic schedulers, and planner-auditor systems [6,7]. Figure 1 shows the resulting structure.

Generative modeling has become a central direction in artificial intelligence, with representative frameworks including variational autoencoders (VAEs) [8], energy-based models (EBMs) [9], generative adversarial networks (GANs) [10], normalizing flows (NFs) [11], and diffusion models (DMs) [1]. In image generation, diffusion models have become especially influential because they offer high-fidelity synthesis, strong mode coverage, and comparatively stable training behavior. Unlike GANs, which depend on adversarial optimization, diffusion models explicitly learn a reverse denoising process from Gaussian noise to data, reducing mode-collapse risks while supporting high-resolution synthesis, text guidance, and structured control.

This section reviews the theoretical foundations and methodological evolution of diffusion models for image generation, establishing the basis for later discussion of collaborative control.

1.1. Core Theory and Methodological Evolution of Diffusion Models

The theoretical foundation of diffusion models is rooted in stochastic processes and Brownian motion. Brownian motion describes the evolution of particles under random perturbations and can be formulated as a stochastic differential equation (SDE):

$$dX_t = \mu(X_t, t) dt + \sigma(X_t, t) dW_t \quad (1)$$

where X_t is the random state at continuous time t , $\mu(X_t, t)$ is its drift, and $\sigma(X_t, t)$ is the diffusion amplitude. The increment dt denotes an infinitesimal time interval, and W_t is standard Brownian motion. Diffusion models adopt this idea by perturbing data into a tractable noise distribution and learning the reverse process.

A related mathematical basis is the diffusion equation in thermodynamics,

$$\frac{\partial u}{\partial t} = D \nabla^2 u \quad (2)$$

where $u = u(\mathbf{r}, t)$ is a scalar field at spatial coordinate $\mathbf{r}$ and time t . The coefficient D controls diffusion speed, and ∇^2 is the spatial Laplacian. This equation describes how local differences are progressively smoothed over time. In diffusion models, the forward process similarly erases data structure, while the reverse process learns score-based corrections.

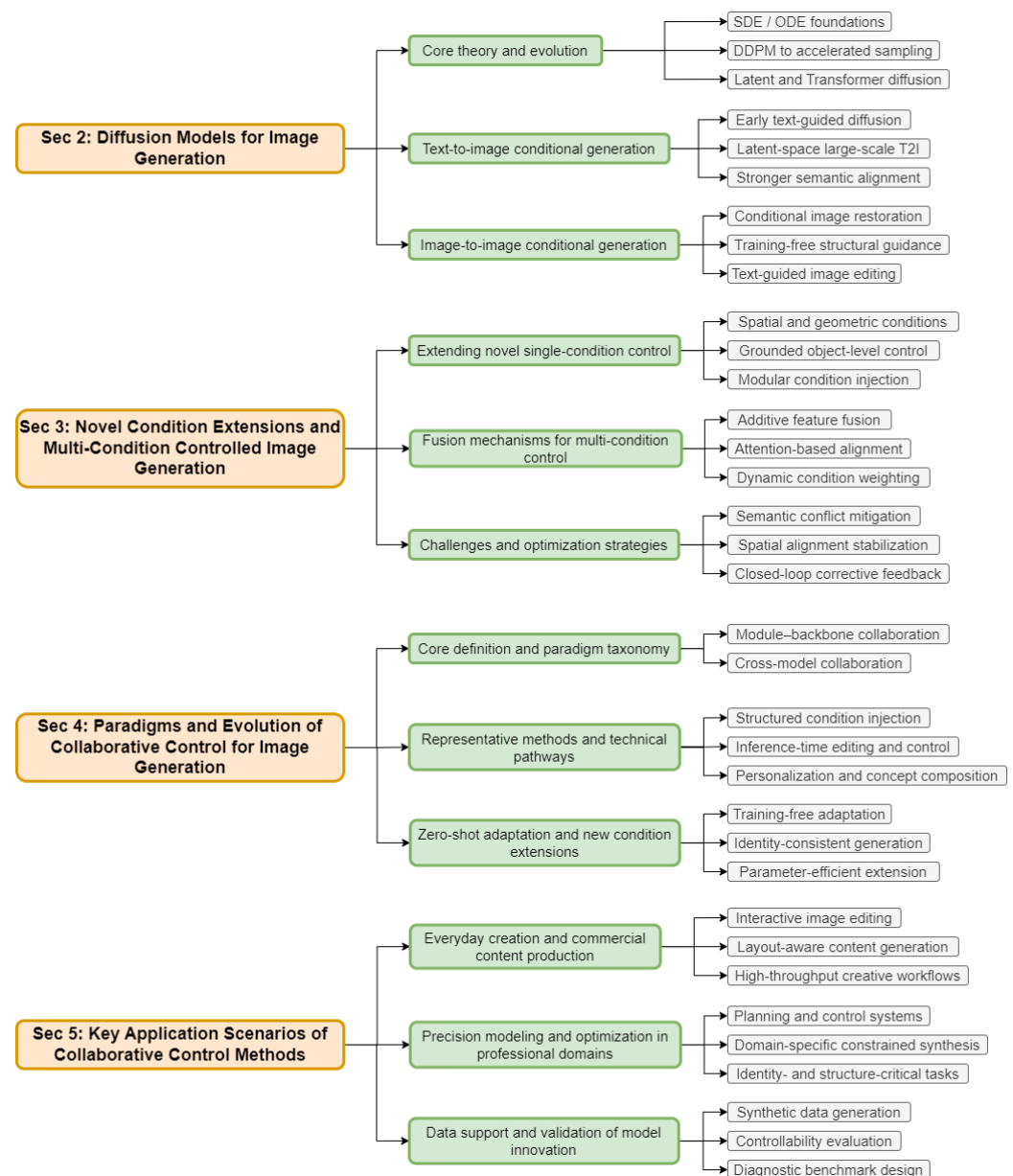


Figure 1. Overall structure of this survey. This figure presents the main sections of the paper and their hierarchical thematic organization.

The idea of diffusion-based generation predates modern DDPMs. Sohl-Dickstein et al. proposed a nonequilibrium thermodynamics perspective in which data are gradually transformed into a simple distribution and then generated by learning the inverse process [12]. This formulation decomposes a difficult high-dimensional generation problem into a sequence of local transitions, providing a natural basis for later denoising, guidance, and conditional-control mechanisms.

Diffusion models are likelihood-based generative models composed of a forward noising process and a reverse denoising process. Song et al. showed that discrete diffusion models and continuous-time stochastic processes can be unified under score-based generative modeling, where the central learning target is the score, i.e., the gradient of the log-density [3].

Ho et al. introduced Denoising Diffusion Probabilistic Models (DDPMs), which formulate diffusion as a discrete-time Markov chain [1]. Here, x_0 is a clean sample and x_t

is its noisy state at discrete timestep $t \in \{1, \dots, T\}$. The forward kernel q gradually adds Gaussian noise:

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I}), \quad \beta_t \in (0, 1) \quad (3)$$

where $\mathcal{N}(\cdot; m, V)$ denotes a Gaussian distribution with mean m and covariance V . The identity matrix $\mathbf{I}$ matches the dimension of x_t , and β_t is the variance schedule. Defining $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, the marginal distribution conditioned on x_0 has the closed form

$$q(x_t | x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) \mathbf{I}) \quad (4)$$

which can be reparameterized using independent standard Gaussian noise ε as

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \mathbf{I}) \quad (5)$$

The learned reverse kernel p_θ , with parameters θ , is

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (6)$$

where $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ are the predicted reverse mean and covariance. The model is trained by maximizing a variational lower bound on the data likelihood. In practice, DDPMs commonly use the simplified noise-prediction objective:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t, x_0, \varepsilon} [\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2] \quad (7)$$

where $\mathbb{E}_{t, x_0, \varepsilon}$ averages over sampled timesteps, data, and noise. The network $\varepsilon_\theta(x_t, t)$ predicts the added noise, and $\|\cdot\|_2$ is the Euclidean norm. This objective is stable, efficient, and widely adopted.

From the continuous-time perspective, discrete diffusion models can be viewed as time discretizations of SDEs. Under the variance-preserving (VP) formulation, the forward diffusion process is

$$dx = -\frac{1}{2}\beta(t)x dt + \sqrt{\beta(t)} dW_t \quad (8)$$

where $x = x(t)$ is the continuous-time state and $\beta(t) \geq 0$ is its noise-rate schedule. The corresponding reverse-time SDE is

$$dx = \left[-\frac{1}{2}\beta(t)x - \beta(t)\nabla_x \log p_t(x) \right] dt + \sqrt{\beta(t)} d\bar{W}_t \quad (9)$$

where $p_t(x)$ is the marginal density of state x at time t , and $\nabla_x \log p_t(x)$ is its score. The process $\bar{W}_t$ denotes reverse-time Brownian motion. Once the score is estimated across noise levels, samples can be generated by solving this reverse-time dynamics [3].

The probability flow ordinary differential equation (probability flow ODE) provides a deterministic counterpart to the reverse-time SDE. Under the VP parameterization with $g(t)^2 = \beta(t)$, it is written as

$$\frac{dx}{dt} = -\frac{1}{2}\beta(t)x - \frac{1}{2}\beta(t)\nabla_x \log p_t(x) \quad (10)$$

The reverse-time SDE and the probability flow ODE share the same marginal distributions when they use the same score field. This SDE/ODE view explains the connection between stochastic sampling, deterministic sampling, and efficient numerical solvers [3,13].

Under Gaussian forward kernels, the DDPM noise-prediction objective is closely related to score learning. The noise predictor $\varepsilon_\theta(x_t, t)$ and the score $\nabla_x \log p_t(x_t)$ can be

converted through known coefficients. Thus, noise prediction provides an indirect estimate of the score field and supports reverse SDE and probability flow ODE sampling.

Several later studies further unified diffusion training and sampling design. Karras et al. showed that noise schedules, scaling, and preconditioning can be organized within a continuous noise-parameterization framework [14]. Flow Matching formulates generation as velocity-field regression along probability paths [15], while Rectified Flow straightens transport trajectories to enable efficient deterministic generation [16]. These methods connect diffusion models with continuous normalizing flows and provide a broader dynamical-systems view of generation.

Beyond DDPMs, NCSN methods use noise-conditional networks to estimate score functions across multiple noise levels [17]. Adversarial score matching improves perceptual quality through adversarial objectives [18], and Schrödinger Bridge formulations relate score-based generation to optimal transport [19]. Together, these works position diffusion models as programmable stochastic dynamical systems rather than only black-box generators.

Sampling efficiency has been a central direction in diffusion research. DDIM constructs a non-Markovian process while preserving the DDPM training objective, enabling faster deterministic or semi-deterministic sampling [20]. The improved DDPM refines noise schedules, objectives, and parameterizations to improve sample quality and likelihood [21]. These methods show that the same trained model can support different reverse trajectories, making sampler design a key factor in the speed–quality trade-off.

ODE-based samplers further accelerate generation. DPM-Solver treats sampling as numerical solution of the diffusion ODE and uses high-order methods to reduce function evaluations [13]. Progressive Distillation compresses multi-step teacher sampling into few-step student models [22]. UniPC introduces a unified predictor–corrector framework for high-order correction without additional network evaluations [23]. Optimized Time Steps further show that non-uniform time discretization can reduce error accumulation under small sampling budgets [24].

Diffusion models have also become competitive with GANs in high-fidelity synthesis and have surpassed them in several benchmark settings. Dhariwal and Nichol demonstrated that diffusion models with improved architectures and classifier guidance can achieve strong image-generation quality [2]. For controllability, Semantic Diffusion Guidance injects semantic gradients into pretrained diffusion models without retraining [25]. Classifier-free guidance (CFG) combines conditional and unconditional scores to balance alignment and diversity [26]. S-CFG further extends CFG by assigning different guidance strengths to semantic regions, improving prompt following and local control [27].

Latent Diffusion Models (LDMs) improve scalability by performing diffusion in the latent space of a pretrained VAE rather than directly in pixel space [28]. This reduces computational cost while preserving visual quality. An LDM also uses cross-attention to inject text and other conditions into the denoising network, establishing a widely adopted interface for conditional diffusion and text-to-image generation.

Architectural evolution has further expanded diffusion-model capability. U-Net remains a dominant backbone because of its multi-scale representation and skip connections [29]. Diffusion Transformer (DiT) replaces convolutional U-Net with Transformer blocks to improve global modeling and scalability [30]. Hybrid architectures combine convolutional locality with Transformer-based long-range modeling [31]. MaskDiT accelerates training through masked token reconstruction [32], while MDT and MDTv2 improve masked latent modeling and contextual learning among image tokens [33]. FiT supports flexible resolutions and aspect ratios through improved positional encoding [34].

For multimodal and structured generation, MMDiT concatenates text and image tokens and applies unified attention to strengthen cross-modal interaction [35]. Graphusion

extends diffusion modeling to graph-structured data through latent graph generation [36]. D2iT further addresses fixed-compression limitations by using dynamic multi-granularity compression and adaptive allocation for information-dense regions [37].

Overall, diffusion models have evolved from stochastic noising and denoising processes into a broad family of score-based, ODE/SDE-based, latent-space, and Transformer-based generative frameworks. Despite these advances, precise controllable generation still faces challenges in semantic alignment, few-shot robustness, inference efficiency, and multi-condition consistency. These limitations motivate the subsequent discussion of collaborative control mechanisms for precise image generation.

1.2. Text-to-Image Conditional Image Generation

Text-to-image (T2I) conditional image generation refers to the task of synthesizing visually coherent, semantically aligned, and high-fidelity images from natural language descriptions. This task can be formalized as learning an accurate and interpretable mapping between textual semantics, including object attributes, spatial relations, scene context, and style requirements, and visual pixel representations [4]. In contrast to unconditional image generation, which aims to produce realistic images without explicit constraints, T2I generation requires models to capture subtle linguistic distinctions, resolve semantic ambiguity, and translate abstract textual concepts into concrete visual details. From a controllable generation perspective, this boundary also reflects an important engineering reality: textual conditions are inherently weakly structured and lack a deterministic syntax that directly grounds to pixels or layout. Consequently, models rely on vision–language alignment representations, such as contrastively learned joint embeddings, to map open-vocabulary semantic constraints into local evidence that can be progressively enforced along the denoising trajectory [38].

Before diffusion-based T2I approaches became dominant, research largely followed two directions. One direction focused on autoregressive generation over discrete visual tokens. A representative example is DALL-E, which builds on VQ-VAE-style discrete codes and models image generation using Transformers [39] in a discrete sequence space, enabling limited semantic composition [40]. The other direction centered on adversarial learning, where conditional GANs mapped text descriptions to images through supervised generator–discriminator training, but often suffered from instability and limited semantic alignment, particularly for high-resolution detail and long textual prompts. As diffusion models demonstrated advantages in fidelity and training stability, attention shifted toward a unified framework of conditional diffusion with explicit alignment mechanisms. The key advance was not only improved image quality, but also the application of conditioning throughout a multi-step denoising process, allowing textual constraints to influence generation in a staged manner across noise scales. Early denoising stages primarily determine global composition and semantic layout, whereas later stages focus on texture synthesis and fine-detail consistency. This property provides a natural basis for staged conditioning and mixture-of-experts designs [41].

Major progress along the diffusion-based T2I trajectory can be attributed to advances in three areas: stronger text representations, more stable conditional guidance, and more efficient generation carriers. GLIDE [42] represents an early text-conditional diffusion framework that integrates textual information into the diffusion process and systematically studies guidance strategies that balance photorealism and text alignment. Its introduction of classifier-free guidance (CFG) established a general mechanism for controllable sampling under free-form prompts. Imagen [43] further highlighted the importance of text encoders for image–text alignment. Rather than jointly training a text encoder on paired data, Imagen employs a pretrained and frozen language model such as T5 [44], leveraging large-scale

text-only corpora to improve semantic representation and enhance alignment and fidelity through CFG during generation.

From an engineering scalability perspective, Latent Diffusion Models (LDMs) and Stable Diffusion [28] move diffusion into a compact latent space, significantly reducing computational cost through a pretrained autoencoder. These models employ cross-attention as the primary conditional interface, allowing text tokens and visual latents to align during denoising. This design improves the practicality of high-resolution generation and supports an ecosystem of prompt engineering techniques, including negative prompts and guidance weight scheduling, which enable finer control over the trade-off among text constraint strength, image quality, and diversity. Building on this foundation, eDiff-I analyzes differences in how diffusion sampling depends on text at early versus late stages and improves alignment through staged expert denoisers without increasing inference cost [41]. For larger-scale backbones and stronger performance, SDXL enhances detail rendering and aesthetic quality through improved architectures and conditional training strategies [45]. Systems such as DALL-E 3 emphasize more robust text understanding and prompt following, improving reliability under complex semantics, typography, and long prompts [46]. In large-scale industrial settings, PixArt- α integrates cross-attention into DiT and adopts staged training and data strategies to balance cost and quality [47], while PixArt- Σ extends this direction through weak-to-strong training and efficiency improvements for higher-resolution synthesis [48]. Hunyuan-DiT focuses on fine-grained bilingual understanding, multi-resolution modeling, and data pipeline design to strengthen large-scale T2I generation [49]. In cascaded diffusion systems, DeepFloyd IF employs a frozen text encoder together with a multi-stage pixel-space diffusion cascade to distribute high-resolution detail recovery across stages, improving photorealism and scalability while preserving text understanding [43].

Recent foundation models extend this progression from latent U-Nets toward flow-based Transformer systems and unified multimodal interfaces. FLUX.1 uses a large rectified-flow Transformer backbone, and its open-weight FLUX.1-schnell variant supports few-step generation under an Apache-2.0 license [50]. This backbone is important for collaborative control because its multimodal token interactions provide a different injection surface from U-Net residual branches. OmniGen goes further by representing text, source images, and target-generation instructions in one Transformer sequence [51]. It supports text-to-image generation, subject conditioning, and image editing without a separate architecture for each task. Compared with earlier specialized adapters, this unified interface reduces task-specific engineering, but it transfers more responsibility to prompt interpretation and increases backbone-level computation.

OminiControl represents a complementary recent development for DiT-based control [52]. Instead of adding a large parallel branch for each task, it reuses the backbone's VAE encoder and Transformer blocks, concatenates condition and image tokens in one sequence, and adapts positional encoding to spatially aligned and non-aligned conditions. Its reported 0.1% additional parameter requirement illustrates a shift from task-specific control architectures toward compact interfaces that support several forms of image conditioning. Relative to OmniGen's unified generation interface, OminiControl focuses more narrowly on parameter-efficient condition injection into an existing diffusion Transformer.

Despite substantial progress in image quality and open-domain semantic expressiveness, text conditioning continues to exhibit structural limitations. Uncertainty in natural language with respect to spatial layout, object counting, and attribute binding can be amplified during denoising into local mismatches or relational errors. Moreover, cross-attention alignment under long prompts and multi-entity scenes is susceptible to token competition, attention dilution, and semantic leakage, leading to unstable fine-grained text-

image consistency. These limitations motivate subsequent methodological developments in structured condition extensions, multi-condition fusion, and collaborative control. Typical diagnosable failure modes include inconsistent object counts, incorrect attribute transfer among entities, and the emergence of hallucinated local content unrelated to the prompt. Such observations indicate that soft alignment via cross-attention alone is insufficient to provide verifiable structural constraints, thereby motivating the introduction of explicit spatial conditions, region-level constraints, and closed-loop correction mechanisms that translate textual satisfaction into more stable process-level control signals.

1.3. Image-to-Image Conditional Image Generation

Image-to-image (I2I) conditional generation refers to the task of synthesizing a target image x_{tar} given a source image x_{src} as a condition. This task encompasses applications such as image inpainting, style transfer, and super-resolution, and relies on deep generative models to learn the conditional distribution $p(x_{\text{tar}} | x_{\text{src}})$. The objective is to generate visually plausible outputs while preserving structural consistency with the input. Compared with text-to-image generation, I2I conditions operate directly in high-dimensional visual space, requiring a finer balance among content preservation, local editability, and detail consistency. Within the diffusion framework, this balance can be articulated as maintaining condition-consistent low-frequency structure, including layout, contours, and geometry, while allowing high-frequency details, such as texture, material appearance, and noise coherence, to be guided by the generative prior along the sampling trajectory. Different tasks impose different relative emphases on these aspects [53].

Historically, I2I conditional generation was dominated by conditional GANs, which formulate the task as supervised mapping between paired image domains. Pix2Pix introduced one of the earliest widely adopted paired I2I frameworks [54], while CycleGAN removed the requirement for paired data through cycle consistency [55]. MUNIT and DRIT further enabled multimodal outputs by disentangling content and style representations [56,57]. Despite their success, GAN-based approaches exhibit inherent limitations in training stability, mode coverage, and modeling complex conditional distributions, motivating a transition toward diffusion models with explicit probabilistic formulations and stable optimization. A key driver of this transition is controllable reconstruction: diffusion models naturally support gradual recovery from noise and allow conditioning to be injected at different noise scales, aligning well with practical image editing requirements in which structure must be preserved while new details are synthesized [58].

Diffusion models generate images by transforming random noise through a sequence of denoising steps, providing a probabilistic alternative to adversarial learning. In the I2I setting, the central innovation lies in conditioning the reverse diffusion process on the source image so that the output remains consistent with the input in both semantics and structure. Work such as Palette demonstrates that diffusion models achieve strong performance on tasks including colorization, inpainting, and reconstruction, and can unify multiple I2I tasks within a single framework without designing task-specific architectures [59]. Beyond unified multitask learning, training-free I2I guidance represents another important contribution of the diffusion paradigm. ILVR injects structural information from a reference image into an unconditional diffusion model by aligning or replacing low-frequency components during sampling, enabling super-resolution, cross-domain translation, and sketch- or scribble-guided control without additional training [53].

Latent diffusion further extends these benefits by performing diffusion in a compressed latent space rather than pixel space. This approach substantially reduces computational cost while retaining high-quality generation and enables injection of I2I conditions via cross-attention or feature fusion mechanisms [28]. These advances are particularly important for

super-resolution, which requires faithful reconstruction of both global structure and local high-frequency detail. SR3 represents an early diffusion-based super-resolution method that progressively restores high-frequency information over multiple denoising steps, producing more natural textures under large upscaling factors and enabling cascaded pipelines that combine base models with super-resolution refinement [60].

For image restoration tasks, recent work applies diffusion models to inpainting, denoising, and deblurring. DiffIR combines a dynamic restoration Transformer with a compact diffusion prior to improve stability and reduce the number of denoising steps required for high-quality restoration [61]. In super-resolution, ResShift introduces a residual shift mechanism into the diffusion process to accelerate sampling while preserving high-frequency detail under fewer steps [62]. For free-form inpainting, RePaint modifies only the reverse sampling procedure by repeatedly reinjecting observed pixels in unmasked regions, allowing an unconditional DDPM to adapt to extreme masks and diverse missing patterns without retraining [63]. This approach illustrates the flexibility of sampling-time constraints for I2I tasks.

The interaction between input image features and intermediate model representations is a critical factor in the success of diffusion-based I2I methods. Cross-attention, feature concatenation, and multi-scale fusion are commonly used to integrate input conditions into the denoising process, enabling preservation of global coherence while progressively restoring fine detail. In inpainting and super-resolution, structural cues such as segmentation masks, edges, and depth maps are particularly important for suppressing undesired content drift and enhancing structural consistency. From a control-interface perspective, I2I conditions can function as hard observations, such as unmasked pixels or low-resolution inputs that are enforced during sampling, or as soft conditions, such as reference style or semantic edit directions that influence denoising through alignment losses or guidance gradients. The combined use of these two condition types forms a basic paradigm for subsequent collaborative control, in which hard constraints and soft guidance operate in parallel [64].

In summary, diffusion models provide a stable, flexible, and high-quality probabilistic generation framework that shifts I2I conditional generation away from adversarial learning toward explicit conditional consistency constraints and controllable denoising trajectories. This shift establishes a solid foundation for later developments in multi-condition fusion and collaborative control. At the same time, text-guided I2I editing has emerged as a bridge between T2I and I2I paradigms by using real images as anchors for semantic manipulation. Methods such as DiffusionCLIP [58] and Blended Diffusion [64] leverage vision–language models to translate textual constraints into optimization signals, enabling robust zero-shot editing and improved background preservation through global semantic control and region-level editing, respectively.

2. Condition Extensions and Multi-Condition Fusion for Controlled Generation

The central problem of conditional diffusion is to align an external condition c with intermediate denoising representations so that generation satisfies the constraint while preserving the learned prior. In a general form, the reverse transition becomes

$$p_{\theta}(x_{t-1} \mid x_t, c) = \mathcal{N}(x_{t-1}; \mu_{\theta}(x_t, t, c), \Sigma_{\theta}(x_t, t, c)), \quad (11)$$

where p_{θ} is the learned reverse kernel and θ denotes its parameters. The variable c may represent text, masks, depth, poses, reference images, or external knowledge. The functions μ_{θ} and Σ_{θ} are the condition-dependent reverse mean and covariance. Condition injection enters through three interfaces: the input state, intermediate features, or attention. These

interfaces determine control granularity, interpretability, and the location where conflicts emerge under composition [39,65].

Text-to-image models provide open-ended generation but cannot reliably specify complex layouts, strict geometry, or editable regions through language alone [5,7]. Condition extension and multi-condition fusion therefore upgrade user intent from weak textual hints to executable signals, such as geometry, topology, region semantics, and interaction trajectories, reducing prompt engineering and uncertainty [63].

2.1. Novel Single-Condition Control Extensions

Novel single-condition control introduces structured or observable constraints that compensate for text in spatial layout and geometry. Conditions can be dense pixel-level signals, such as edges, depth, and segmentation; sparse object-level signals, such as bounding boxes, keypoints, and scene graphs; or interactive process signals, such as drag trajectories and brush hints. These forms determine the condition encoder, injection location, and conflict-resolution strategy [53]. Instead of relying on a weak prompt-to-pixel path, such methods create a stronger chain from condition space to intermediate representations and then to denoising updates, reducing object misplacement, relation confusion, and structural drift [6]. Many conditions directly encode target spatial structure and geometric relations, including sketches, edges, depth, semantic segmentation, bounding boxes, and scene graphs. Yang et al. proposed a diffusion-based scene-graph-to-image framework that learns graph-structured representations aligned with real images at both global and local levels and uses them as conditions for training a latent diffusion model, improving adherence to scene-graph constraints [66]. For sketch-like spatial conditions, Voynov et al. proposed an inference-time guidance paradigm based on a lightweight Latent Guidance Predictor that maps intermediate latent features in a diffusion model to spatial map predictions and injects a spatial consistency loss during sampling to progressively align generation with a target sketch. This approach enables spatial control without training a dedicated encoder for each condition type. It also unifies supervision across spatial maps, including sketches, edges, and segmentations, by expressing them as differentiable constraints that can be evaluated during inference. Rather than relying on end-to-end encoder injection, the sampling trajectory is repeatedly corrected by computable consistency objectives, which naturally supports reuse across diverse spatial conditions [67]. Within training-free spatial control, ILVR provides an instructive reference by aligning low-frequency structure during sampling. This result indicates that, even without explicit ControlNet or adapter branches, structural adherence can be enforced through frequency-domain and scale-based operations that act directly on the denoising trajectory, suggesting transferable designs that combine lightweight guidance with structural conditions.

Another class of conditions emphasizes controllable correspondences between textual entities and image regions, often specified as paired localization and semantics in the form of bounding boxes or segments together with captions or phrases. Li et al. proposed GLIGEN, which injects grounding conditions through newly introduced trainable layers with gating while keeping the backbone of a pretrained text-to-image diffusion model frozen. This design enables joint conditional generation with captions and spatial constraints, such as bounding boxes, and maintains strong generalization and zero-shot grounding for open-set concepts and novel spatial configurations [68]. This line of work upgrades object-level controllability from prompt engineering to verifiable spatial constraints. Once object-to-region correspondences are specified explicitly, attribute binding errors in multi-entity scenes can be reduced, providing stable anchors for more complex coordination, such as combining region text with segmentation or depth. At an implementation level, GLIGEN remains effective with a frozen backbone because grounding is injected as

an additional gated control channel at feature interaction sites, allowing spatial anchors to influence update directions of denoising residuals rather than acting only as indirect hints on the text side [68].

In practical settings, high-precision dense geometric conditions, such as accurate depth maps, are often difficult to obtain. This motivates methods that use relaxed geometric constraints as interactive alternatives. Bhat et al. addressed this bottleneck by generalizing depth conditions from dense maps to weaker geometric constraints, such as boundaries and 3D boxes, and by emphasizing generalization of geometric control and interactive workflows under weak conditions. A representative contribution of LooseControl is the reinterpretation of depth conditioning as an editable geometric scaffold interface. Users provide low-cost constraints, such as room boundaries or 3D boxes, and the model synthesizes unspecified details while maintaining global perspective consistency, which aligns with iterative content production workflows that establish structure first and refine details later [69]. This trend shifts part of the controllability burden from the data side, where annotations are expensive, to the interaction side, where inputs are lower cost. It trades stronger geometric guarantees for sufficiently useful layout consistency and better matches practical processes of rapid iteration and progressive correction [69]. Methodologically, it pushes single-condition control from pursuing maximal constraint strength toward a tunable trade-off between constraint strength and degrees of freedom. When a condition is incomplete, the system typically requires explicit fault-tolerance mechanisms, such as constraining only low-frequency structure or activating the condition only at selected noise stages, to avoid artifacts or structural breaks that may arise from hard injection of imperfect constraints.

At the level of injection interfaces, ControlNet [6] and T2I-Adapter [7] represent two typical paths, namely a frozen backbone with a dedicated control branch and a lightweight adapter, respectively. The former achieves strong control through a stable branching structure around a frozen backbone, whereas the latter supports lower-cost condition extension and compositional deployment through plug-in adapters. In practical systems, the distinction is often reflected in a three-way trade-off among control strength, model footprint, and compositional flexibility. ControlNet favors strong constraints with reliable empirical performance, whereas adapters favor lightweight composition and rapid iteration. This trade-off directly affects the complexity of conflict management in multi-condition systems and the scheduling policies used at inference time. A common product-oriented strategy is to use a strong ControlNet-style module to lock a critical spatial scaffold, such as pose, segmentation, or depth, and to use lighter components, including adapters and attention manipulation, to introduce secondary semantics or style preferences. This reduces the decision pressure that arises when multiple strong constraints are stacked simultaneously.

2.2. Fusion Mechanisms for Multi-Condition Control

Multi-condition control integrates two or more signals, such as text and edges, segmentation and pose, or text and depth. Its central challenge is spatial and semantic conflict among conditions. Mainstream fusion strategies include parallel additive fusion, cross-attention fusion, and dynamic weight fusion. A compact representation is

$$h'_l = h_l + \sum_{i=1}^N w_i(t, l) G_i(c_i), \quad (12)$$

where h_l and h'_l are the pre-fusion and post-fusion features at layer l . The set contains N conditions, c_i is condition i , and G_i is its encoder. The schedule $w_i(t, l)$ controls condition

strength at timestep t and layer l . Static fusion fixes these weights, whereas dynamic fusion adapts them to the temporal roles of conditions [70]. Conflict mediation can be expressed as

$$\min_{\{w_i(t,l)\}} \sum_{i=1}^N \alpha_i \mathcal{L}_i(\hat{x}_0, c_i) + \beta \sum_{i < j} \Omega_{ij}(c_i, c_j, w_i, w_j), \quad (13)$$

where $\hat{x}_0$ is the clean-sample estimate and $\mathcal{L}_i(\hat{x}_0, c_i)$ measures adherence to condition i . The coefficient $\alpha_i \geq 0$ specifies its priority, while $\beta \geq 0$ controls the pairwise conflict penalty Ω_{ij} . This penalty measures incompatibility or redundancy between conditions i and j under their schedules. The formulation turns output conflicts into process-level decisions [14]. Parallel additive fusion established the first scalable path for structured control. ControlNet transforms each condition through a dedicated branch and adds its residuals to matched U-Net features [6]. This design provides strong spatial adherence and preserves the pretrained backbone, but parameter and memory costs grow with the number of branches. T2I-Adapter and Uni-ControlNet respond to this limitation by compressing control into lightweight or shared adapters [7,71]. They improve storage and compositional flexibility, although small adapters may provide weaker control when geometry is strict or inputs conflict.

Attention-based methods address a different limitation: additive features do not explicitly bind semantic entities to regions. Make-A-Scene couples segmentation with text through cross-attention, while SpaText constructs spatial text representations for open-vocabulary regions [72,73]. These methods improve attribute binding and local semantics because region tokens interact directly with visual tokens. Their cost grows with region density, and their accuracy depends on the quality of masks or region descriptions. Thus, attention fusion is preferable when semantic correspondence is primary, whereas feature branches remain stronger for dense geometry.

Inference-time composition avoids training a new joint controller. Composable Diffusion combines conditional scores as energy terms and generalizes to uncommon concept combinations [74]. pix2pix-zero estimates edit directions in text space, while MultiDiffusion couples several regional denoising paths through a shared optimization target [75,76]. These methods are flexible and reusable, but they trade parameter efficiency for repeated denoising, sensitivity to guidance scales, or weaker hard spatial guarantees.

Unified multi-condition models attempt to resolve these trade-offs at the policy level. Composer trains on decomposed factors, including text, depth, sketches, and color, so that arbitrary subsets can be combined at inference [77]. Cocktail regulates branch statistics and spatial scope, while DynamicControl evaluates candidate controls before selecting and weighting them [70,78]. This progression moves the field from unconditional summation toward explicit arbitration. It also reveals a remaining limitation: an arbitration policy can only be as reliable as its conflict detector and training coverage.

Across these families, performance differences follow from where and when a constraint acts. Dense feature branches usually outperform lightweight adapters on strict edges, depth, and pose because they influence several resolutions. Region attention is more effective for object–attribute binding because it provides explicit correspondences. Multi-path inference handles disjoint regions well but scales with the number of paths. Dynamic scheduling is most useful when inputs are redundant or inconsistent, although it adds evaluator and policy cost. The relevant comparison is therefore not a single quality score, but a Pareto balance among constraint fidelity, composability, robustness, and compute.

Table 1 compares representative fusion methods from the perspectives of condition composition, integration interface, adaptation requirement, and practical trade-offs.

Table 1. Representative multi-condition fusion strategies for collaborative control in diffusion models.

Method	Conditions	Interface	Fusion/Arbitration	Main Trade-Off
ControlNet [6]	text + spatial controls	feature branch	additive branch injection	strong control; limited explicit conflict arbitration
SpaText [73]	global text + region text + segmentation	attention	region-wise text guidance	fine-grained semantics; costly with dense regions
Composable Diffusion [74]	multiple textual concepts	sampling score	compositional score fusion	flexible without re-training; weak spatial grounding
MultiDiffusion [76]	text + regional constraints	trajectory/sampling	coupled regional diffusion paths	strong layout control; higher inference cost
Composer [77]	text, depth, sketch, color	feature injection	unified factor composition	broad coverage; inconsistent inputs remain difficult
Uni-ControlNet [71]	local + global controls	adapters	unified local/global adapters	scalable interface; variable control strength
Cocktail [78]	multimodal regional controls	feature + sampling	ControlNorm and spatial guidance	robust multimodal fusion; added tuning burden
Dynamic Control [70]	multiple candidate conditions	scheduling	adaptive selection and weighting	handles redundancy/conflict; extra policy cost

Table 2 shows that collaborative control methods optimize different targets: ControlNet and T2I-Adapter emphasize structural or semantic adherence, GLIGEN combines quality with object-level grounding, MultiDiffusion stresses regional composition, and DynamicControl or UniCombine evaluate heterogeneous control with metrics such as Canny F1, HED SSIM, OpenPose mAP, depth RMSE, mIoU, CLIP-I, DINO, and CLIP-T. Thus, multi-condition fusion should be assessed jointly by fidelity, condition adherence, structural alignment, subject consistency, and computational overhead.

Table 2. Quantitative comparison of representative multi-condition fusion and control methods.

Method	Backbone	Task/Dataset	Metrics	Reported Results
ControlNet [6]	Stable Diffusion	ADE20K segmentation; sketch study	FID, CLIP, IoU, human rank	FID 15.27; CLIP 0.26; IoU 0.35 ± 0.14 ; human quality/fidelity 4.22/4.28.
T2I-Adapter [7]	SD 1.4	COCO text + segmentation/sketch	FID, CLIP	Seg.: FID 16.78, CLIP 0.2652; sketch: FID 17.36, CLIP 0.2666.
GLIGEN [68]	LDM/SD	COCO grounded generation	FID, YOLO AP	COCO2014D: FID 5.61, AP/AP50/AP75 24.0/42.2/24.1; COCO2017 FID 21.04.
MultiDiffusion [76]	Stable Diffusion	text-to-panorama	FID, CLIP, aesthetic	FID 10.3 ± 4.8 ; CLIP 0.27; aesthetic 6.36.
DynamicControl [70]	SD 1.5	MultiGen-20M, ADE20K, COCO-Stuff	F1, SSIM, mAP, RMSE, mIoU	Canny F1 39.26; HED SSIM 0.8376; OpenPose mAP 82.63; depth RMSE 23.21.
UniCombine [79]	FLUX.1-schnell	multi-spatial and subject control	FID, F1, MSE, CLIP-I/DINO/T	Multi-spatial: FID 6.82, F1 0.64, CLIP-T 33.45; subject insertion: FID 4.55.

2.3. Challenges and Optimization of Multi-Condition Control

Multi-condition control expands the controllable space but introduces structural challenges: semantic inconsistency, redundant signals, cross-resolution alignment errors, long-tail condition combinations, and expensive paired annotations. These failures appear as concept drift, attribute binding errors, boundary misalignment, local artifacts, or suppression of one condition by another. Recent optimization strategies progress from feedforward

stabilization to semantic disentanglement, spatial alignment, and closed-loop feedback. A general training or inference objective can be written as

$$\mathcal{J} = \mathbb{E}_{t, x_0, \varepsilon} \left[\|\varepsilon - \varepsilon_\theta(x_t, t, \mathcal{C})\|_2^2 + \sum_{i=1}^N \lambda_i(t) \mathcal{L}_{\text{ctrl}}^{(i)}(\hat{x}_0^{(t)}, c_i) + \gamma(t) \mathcal{L}_{\text{cons}}(\hat{x}_0^{(t)}, \mathcal{C}) \right], \quad (14)$$

where $\mathcal{J}$ is the total training or sampling objective and $\mathcal{C} = \{c_i\}_{i=1}^N$ is the condition set. The estimate $\hat{x}_0^{(t)}$ is recovered from x_t at timestep t . The loss $\mathcal{L}_{\text{ctrl}}^{(i)}$ measures adherence to condition i , and $\mathcal{L}_{\text{cons}}$ measures cross-condition or cycle consistency. Crucially, $\lambda_i(t) \geq 0$ and $\gamma(t) \geq 0$ are timestep-dependent schedules rather than fixed constants. Piecewise, monotone, or learned schedules can prioritize layout at high noise, then identity, semantics, and texture at lower noise. This staged objective represents the denoising properties discussed throughout the survey and clarifies why unstable schedules amplify spatial or semantic errors [70,78]. At the feedforward level, differences in modality and inconsistent injection scales are major sources of instability, motivating normalization and injection designs that reduce mismatches in gradient scales and feature statistics. Cocktail aligns multimodal control signals with gControlNet and uses ControlNorm to regulate feature statistics and influence strength of different condition branches at injection time. Together with spatially guided sampling, this approach reduces unintended changes outside specified regions and mitigates instability and condition distortion under multimodal composition [78]. Complementary to this direction, Uni-ControlNet reduces the cost of multi-condition control through architectural organization. Under a frozen pretrained diffusion backbone, it introduces two classes of adapters, local control and global control, such that model size and fine-tuning cost do not scale linearly with the number of condition types, and it emphasizes composable multi-condition control within a single model [71]. Beyond deployment efficiency, a unified interface also standardizes the action space for condition selection and scheduling. When different conditions enter the denoising chain through isomorphic interfaces, transferable dynamic weighting policies become easier to learn. For more general composition, UniCombine proposes a unified architectural paradigm based on DiT [30] and conditional attention, aiming to maintain consistent alignment under arbitrary condition subsets. It also constructs datasets that compose subject conditions with spatial conditions to mitigate instability caused by limited coverage in training [79]. From a systems perspective, these methods pursue sublinear cost scaling with respect to the number of supported conditions. If each new condition requires an additional ControlNet branch, storage and deployment costs increase linearly. Unified adapters and composable interfaces therefore become important for engineering multi-condition collaboration. This also aligns with the recent shift from adding isolated new conditions toward ensuring stable behavior under arbitrary combinations, which better matches open-world inputs in practical applications.

At the level of semantic consistency, multi-condition failures often originate from semantic conflicts and representational entanglement. Semantic conflicts occur when global text, region text, and layout or relation constraints disagree, whereas entanglement commonly arises in subject customization or multi-concept injection when identity information becomes contaminated by mutable attributes. To reduce conflicts and improve composability, DC-ControlNet decomposes global control into a hierarchical conditioning system and introduces intra-element and inter-element controllers that separately handle object-internal content or layout signals and object-relational signals, such as occlusion. This reduces misinterpretation and mutual suppression when multiple elements and conditions are applied simultaneously [80]. DynamicControl formulates selection, ordering, and weighting of conditions as a learnable decision problem. It uses condition evaluation

and dynamic multi-control adapters to adaptively select effective signals under multi-condition inputs, reducing the negative influence of redundant or noisy conditions on the generation trajectory [70]. DivControl decomposes ControlNet-style control into reusable base components and condition-specific tailoring modules, aiming to reduce knowledge interference across conditions and improve parameter efficiency for multi-condition extension within a unified framework [81]. For entanglement in subject-driven generation, DisenBooth performs disentangled fine-tuning during denoising by separately learning identity-preserving and identity-irrelevant embeddings and by introducing auxiliary objectives, including weak denoising and contrastive embedding losses, to reduce identity drift and dominance of irrelevant factors. This improves identity stability and control flexibility when combining subject conditions with pose, background, and text [82]. In training-free multi-concept composition, LoRA-Composer mitigates concept vanishing and concept confusion by introducing concept injection and isolation constraints together with latent re-initialization, reducing suppression and interference when multiple LoRA modules are combined [83]. Concept Sliders parameterize concept control as continuous sliders along low-rank directions, enabling interpretable and smoothly adjustable concept strengths while reducing interference with other attributes [84].

For spatial alignment, multi-object and multi-condition composition can amplify misalignment errors of structural conditions across resolutions and noise stages, motivating mechanistic constraints that reduce dependence on heuristic tuning. LayoutDiffusion addresses layout-to-image generation by unifying layout and image representations and explicitly modeling multi-object relations through a Layout Fusion Module and Object-aware Cross Attention. This improves position sensitivity and object-level consistency in multi-object scenes and reduces drift under complex layouts [85]. Under realistic inputs where visual control and text prompts are not perfectly aligned, Mask-guided Prompt Following distinguishes aligned and misaligned regions using object masks and introduces a Masked ControlNet that targets synthesis in misaligned regions, improving prompt following and robustness to imperfect control signals under multi-condition inputs [86]. In addition, ControlNet++ introduces pixel-level cycle consistency by using reconstruction error from condition to generation back to condition as an optimization signal and incorporates reward models to provide consistency feedback at relatively low cost. This provides a general mechanism for improving controllability when spatial alignment cannot be achieved in a single forward pass [87]. When feedforward fusion and structured modeling remain insufficient for long-tail combinations, closed-loop feedback and cycle consistency can serve as practical mechanisms for stabilization. ControlNet++ demonstrates that controllable diffusion can still deviate from input controls under diverse conditions and introduces cycle consistency as an objective. It further constructs consistency rewards with pretrained discriminative models and uses an efficient reward fine-tuning procedure that avoids the cost of multi-step backpropagation, improving condition alignment without changing the overall framework [87]. CycleNet introduces cycle-consistency regularization for unpaired image manipulation under text-guided diffusion, improving content preservation and translation consistency and demonstrating robustness under limited data and cross-domain shifts [88]. GDCC further exploits the duality between layout-to-image generation and object detection by constructing a generation–detection cycle-consistency framework that converts layout alignment errors into measurable supervision through detection feedback, improving layout adherence and recognizability [89]. LightDiff applies multi-condition diffusion to unpaired low-light enhancement, introduces multi-condition adapters with adaptive weight control, and trains with reward signals derived from perceptual models, illustrating a practical pipeline that combines weak or unpaired data with reward feedback for multi-condition stabilization [90]. Across these closed-loop approaches, a shared idea

is to convert condition satisfaction from implicit subjective judgment into computable external metrics, such as detection, segmentation, similarity measures, or reward models, thereby providing additional supervision for long-tail combinations. The cost is typically increased pipeline complexity and tighter coupling between training objectives and external evaluators [91].

Table 3 further links multi-condition control failures to measurable evaluation targets. Edge, pose, depth, and segmentation controls correspond to different types of failure: weak edge alignment can be reflected by low Canny F1 or HED SSIM; pose inconsistency can be measured by OpenPose mAP; geometric drift can be reflected by depth RMSE; and semantic-region misalignment can be measured by segmentation mIoU. Compared with early adapter- or grounding-based methods, ControlNet-style branch injection usually provides stronger structural controllability, while methods such as Cocktail and ControlNet++ further improve multimodal fusion or consistency feedback. DynamicControl achieves the best reported results in this comparison, suggesting that adaptive condition selection is effective when multiple conditions are redundant, noisy, or mutually conflicting. Therefore, failure-aware evaluation is necessary for collaborative control: a method with high visual fidelity may still fail if it cannot preserve spatial, geometric, or semantic constraints under multi-condition composition.

Table 3. Quantitative comparison of representative methods for mitigating multi-condition control failures.

Method	Main Failure Addressed	Canny F1 ↑	HED SSIM ↑	OpenPose mAP ↑	Depth RMSE ↓	ADE20K mIoU ↑	COCO-Stuff mIoU ↑
T2I-Adapter-SDXL [70]	single-condition structural control with lightweight adaptation	28.03	NR	63.89	39.76	NR	NR
T2I-Adapter-SD1.5 [70]	lightweight adapter-based control under Stable Diffusion 1.5	23.66	NR	60.17	48.40	12.60	NR
GLIGEN-SD1.4 [68,70]	grounding and spatial anchoring for object-level control	26.92	0.5641	69.88	38.82	23.77	NR
Uni-ControlNet [70]	unified local and global condition composition	27.31	0.6912	72.71	40.66	19.39	NR
UniControl [70]	unified controllable visual generation	30.83	0.7967	75.87	39.17	25.45	NR
ControlNet-SD1.5 [6,70]	strong branch-based structural injection	34.66	0.7622	NR	NR	32.56	27.47
Cocktail [70]	multimodal control fusion and spatially guided sampling	35.22	0.8152	78.82	35.90	36.55	29.68
ControlNet++ [70]	consistency-feedback enhanced controllability	37.04	0.8097	NR	28.32	43.64	34.56
DynamicControl [70]	adaptive condition selection and conflict reduction	39.26	0.8376	82.63	23.21	48.56	37.78

Note: NR denotes a metric not reported for that method in the common comparison source. It is not treated as a zero score. All values are transcribed from the same DynamicControl evaluation protocol [70].

3. Collaborative Control Paradigms in Diffusion Models

3.1. Definition and Taxonomy of Collaborative Control Paradigms

Definition 1 (Controllable diffusion). *A diffusion system is controllable when an external condition c changes a specified attribute or region of its output distribution. One prompt, one classifier gradient, one mask, or one ControlNet branch is sufficient for this definition. The mechanism need not coordinate several controls or verify their joint satisfaction.*

Definition 2 (Collaborative control). *Collaborative control is the subset of controllable diffusion in which at least two independently represented conditions, controllers, models, or evaluators affect one generation process. Their influence must be coordinated through explicit fusion, routing, scheduling, spatial partitioning, optimization, or feedback. The system must also expose a joint objective or verification rule for their combined constraints.*

The collaborative reverse update can be expressed as

$$x_{t-1} = \Phi_{\theta}(x_t, t; \mathcal{C}, \mathcal{M}, \pi_t), \quad (15)$$

where x_t and x_{t-1} are consecutive noisy states and Φ_{θ} is the reverse operator with parameters θ . The set $\mathcal{C}$ contains external conditions, $\mathcal{M}$ contains auxiliary modules or models,

and π_t is the timestep-dependent coordination policy. A single-condition method belongs only to controllable diffusion unless it adds another independently acting component and an explicit coordination rule. Thus, LDMs and FLUX are enabling backbones, and a single ControlNet is a controller. They become collaborative systems when combined with other controls, planners, or evaluators under a fusion or feedback policy. This boundary yields three collaborative classes. First, module–backbone collaboration combines several branches or adapters with an explicit fusion rule. ControlNet and T2I-Adapter are building blocks in this class when multiple instances or modalities are composed [6,7]. Second, cross-model collaboration assigns distinct roles to planners, generators, segmenters, or auditors. An LDM alone is only a computational substrate because its autoencoder does not arbitrate a control objective [28]. Third, cross-modal collaboration aligns text, images, regions, and geometry, then schedules their influence across the denoising trajectory. These classes overlap when an external planner routes several multimodal controllers.

Large language models (LLMs) and multimodal large language models make the cross-model class explicit. ELLA uses a timestep-aware connector to align LLM semantics with different denoising stages [92]. Self-correcting LLM-controlled Diffusion Models (SLD) instead close the loop by auditing a candidate and issuing add, delete, move, or modify actions [93]. Recaption, Plan, and Generate (RPG) decomposes a scene into region prompts and executes them through regional diffusion [94]. These methods move language models from prompt encoders to planning or auditing roles.

The translation from verbal feedback to an executable correction can be formalized. At iteration k , an auditor A observes candidate image $I^{(k)}$, prompt p , and constraints $\mathcal{C}$. A parser P converts its verbal error $e^{(k)}$ into structured edits:

$$e^{(k)} = A(I^{(k)}, p, \mathcal{C}), \quad \mathcal{Z}^{(k)} = P(e^{(k)}) = \{(a_j, o_j, b_j, s_j)\}_{j=1}^J, \quad (16)$$

where J is the number of edits. Each tuple contains an action a_j , target object o_j , box or region b_j , and strength s_j . A detector or segmenter converts b_j and o_j into a binary or soft mask $M_j^{(k)}$. This step grounds a language-level diagnosis in image coordinates.

For SLD-like correction, the union mask $M^{(k)}$ can restrict a latent update to diagnosed regions:

$$x_t^{(k+1)} = x_t^{(k)} - \eta_t M^{(k)} \odot \nabla_{x_t} \mathcal{L}_{\text{audit}}(D(\hat{x}_0^{(t)}), \mathcal{Z}^{(k)}), \quad (17)$$

where η_t is the update step, $\odot$ denotes elementwise multiplication, and D decodes the clean estimate $\hat{x}_0^{(t)}$. The differentiable loss $\mathcal{L}_{\text{audit}}$ measures whether the structured edit has been satisfied. Implementations may replace this gradient abstraction with explicit masked add, delete, move, or modify operators.

For RPG-like execution, the parser instead produces regional prompts p_j and masks M_j . Regional denoising predictions are fused as

$$\tilde{\epsilon}_t = \frac{\sum_{j=1}^J M_j \odot \epsilon_\theta(x_t, t, p_j)}{\sum_{j=1}^J M_j + \delta}, \quad (18)$$

where ϵ_θ is the denoiser, $\tilde{\epsilon}_t$ is the fused prediction, and $\delta > 0$ prevents division by zero. Overlaps require normalization or priority rules. The auditor can then re-evaluate the fused image and issue another structured plan. Figure 2 summarizes this transition from open-loop generation to evaluator-guided collaborative correction.

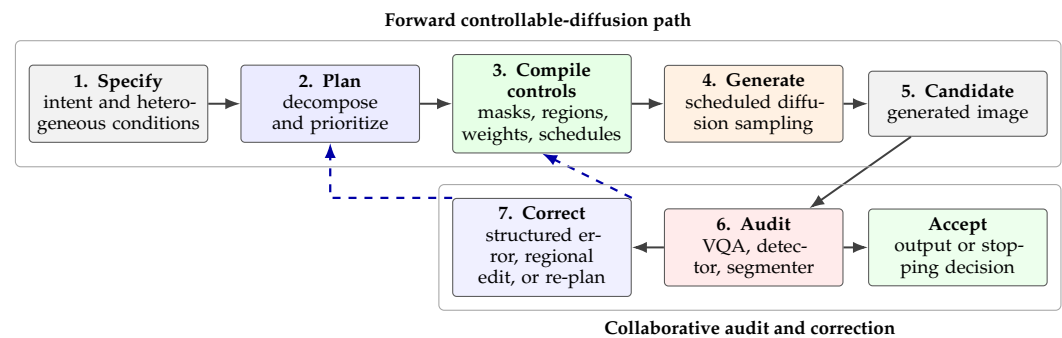


Figure 2. Open-loop generation and closed-loop collaborative correction. The solid path converts intent and heterogeneous conditions into executable controls and a candidate image. Task evaluators either accept the result or return a structured error. Dashed paths distinguish collaborative control by mapping that error to regional correction or re-planning before the next denoising attempt.

Within collaborative control, the methodological spectrum can also be interpreted through two complementary routes, namely control consolidated during training and control overlaid during inference. Training and fine-tuning approaches encode controllability into parameters and the effective data distribution through task objectives. Representative foundation models, such as Imagen [43] and DiT [30], raise the upper bound of controllability by improving backbone expressiveness and enabling scalable training. In contrast, inference-time overlay methods emphasize reuse through modular injection, attention manipulation, guidance functions, or external toolchains without rebuilding the backbone. This lower-cost route is particularly suitable for multi-condition collaboration and rapid extension [43].

3.2. Representative Methods and Technical Pathways

In pathways that coordinate control modules with diffusion backbones, the core goal is to align control signals with key intermediate representations along the denoising process while supporting composable injection. ControlNet freezes the backbone and introduces a trainable control branch. With designs such as zero convolutions, it avoids perturbing the backbone during early training and learns stable mappings from conditions to features [6]. Uni-ControlNet further abstracts control into local and global control adapters, enabling composable superposition of different control modes under a constant adapter size [71]. Cocktail proposes gControlNet and ControlNorm. It first mixes multimodal control signals into a unified embedding and then injects this embedding into the backbone, while using spatially guided sampling to suppress redundant objects outside specified regions [78]. From the perspective of where alignment occurs, these modular collaboration methods do not merely add an auxiliary branch. They select injection interfaces at multi-scale residual blocks, attention blocks, or normalization layers within the U-Net. Injection closer to low-resolution layers tends to constrain global layout and coarse object placement, whereas injection closer to high-resolution layers tends to affect texture, edges, and boundary fidelity. As a result, the allocation of injection strength for a given condition across layers directly determines the trade-off between layout adherence and natural detail, and it influences whether conflicts under multi-condition composition are more likely to appear at a geometric level, such as position and boundaries, or at a semantic level, such as attributes and style [95].

ControlNet++ targets stronger generalization and more robust control. It refines injection mechanisms and training strategies within the ControlNet framework to increase control strength under complex conditions [87]. On the adapter-based collaboration side, T2I-Adapter learns a mapping from condition encodings to backbone interactions using lightweight modules, reducing storage and training overhead while supporting parallel

composition of multiple adapters [7]. IP-Adapter further treats image prompts as plug-and-play adapters injected into text-to-image diffusion models [96]. By decoupling cross-attention to provide an independent channel for image features, it becomes compatible with text prompts and structured control tools. Under the engineering tension between strong control and low-cost extension, the ControlNet family and adapter-based methods can be interpreted as two strategies for channeling conditions. ControlNet provides structured conditions through a parallel feature pathway aligned with the backbone to achieve stronger control reachability, whereas adapters compress conditions into compact plug-in mappings to achieve higher compositional density. When a system must support both structured conditions, such as edges, pose, and segmentation, and example-based conditions, such as reference images and style, an additional attention channel design, such as IP-Adapter [96], can coexist with structural control in parallel and reduce competition in which one condition consumes the attention budget required by another. For DiT backbones, OminiControl extends this efficiency trend by using a shared token sequence and adaptive positional encoding for both aligned spatial controls and non-aligned subject conditions, with only 0.1% reported additional parameters [52]. Beyond these representative methods, recent work has explored designs that trade off efficiency and fidelity under multimodal control, for example by using finer-grained routing and parameter sharing to reduce compute and memory overhead under multi-condition parallelism, thereby pushing composable control from paper-level prototypes toward deployable inference budgets [97].

In cross-model collaboration pathways, a representative direction introduces external alignment or understanding models into the generation loop to improve controllability and interpretability. GLIGEN injects open-vocabulary entities and spatial anchors into the generation process to enable grounded control from text entities to image regions [68]. BoxDiff uses bounding boxes as structural control signals to guide diffusion models to satisfy target layouts [98]. Attend-and-Excite improves controllable generation for key tokens and target objects by explicitly constraining cross-attention focus, mitigating object omission or mismatch caused by textual ambiguity [99]. MultiDiffusion achieves region-level or multi-target collaborative control by binding multiple diffusion paths. It emphasizes parallel sampling for different regions during inference and enforces consistency during fusion, enabling multi-region spatial constraints, local controls such as boxes and segmentation, and global canvas controls such as panoramic generation. The multi-path binding idea exemplified by MultiDiffusion provides a distinct paradigm of cross-module collaboration. Rather than stacking all conditions on a single denoising trajectory, it allows different regions or constraints to follow localized generation trajectories and then merges them into a globally coherent process through shared parameters and consistency coupling. This design is well suited to engineering tasks such as multi-canvas posters, partitioned layouts, long images, and panoramas [76]. In commercial posters, advertising assets, and e-commerce product pages, readable text rendering and controllable layout have become production constraints that are comparable in importance to structural control. TextDiffuser proposes a two-stage paradigm that plans text layouts before diffusion generation. It also constructs large-scale OCR-annotated datasets and benchmarks to evaluate text rendering quality systematically, indicating that text rendering has shifted from an auxiliary capability to a task that can be modeled and evaluated explicitly [100]. TextDiffuser-2 further introduces a large language model for layout planning and dialog-based modification. It makes layout generation and modification an interactive language-driven process and explicitly models line-level positions and content in condition encoding, aligning with iterative typography workflows in practical design [101].

In inference-time overlay control and editing pathways, methods apply control to attention, feature injection points, or sampling trajectories, embedding editing actions into

inference with limited additional cost. Prompt-to-Prompt manipulates cross-attention maps to enable text-only local replacement and editing strength control [102]. InstructPix2Pix constructs instruction-supervised data and trains conditional diffusion models such that natural language editing instructions can be executed in a single forward pass [103]. SDEdit reformulates user guidance as an editing procedure in which noise is added and then removed using diffusion priors, enabling a tunable trade-off between realism and fidelity to the input [104]. DiffEdit automatically generates an editing mask by contrasting model predictions under a source prompt and a target prompt, reducing reliance on manual masks and improving usability for local editing [105]. Region-Aware Diffusion targets region-level text editing and improves semantic consistency while preserving non-edited regions through region-aware entity-level control [106]. MaskSketch introduces sketch structure as sampling-time spatial constraints through a pretrained masked generative Transformer, enabling structure-guided generation without paired supervision or fine-tuning [107]. pix2pix-zero addresses two practical limitations in real image editing, namely incomplete prompt specification and drift in non-target regions. It discovers editing directions in the text embedding space and performs training-free zero-shot image-to-image translation on pretrained text-to-image diffusion models [75]. DragDiffusion converts interactive point dragging into constrained optimization over diffusion features, enabling fine-grained control over shape and position [108]. To integrate these inference-time methods into a unified collaborative control framework, it is helpful to distinguish whether real images must be mapped back into the diffusion domain. One group of methods, such as Prompt-to-Prompt [102] and Attend-and-Excite [99], primarily manipulates attention and denoising dynamics and is best suited to semantic editing of diffusion-generated images. Another group, such as SDEdit [104], DiffEdit [105], and pix2pix-zero [75], explicitly operates on the trajectory of noising and denoising from an input image and emphasizes content preservation and edit directionality. In real image editing, the quality of inversion often determines an upper bound on controllability. Null-text Inversion [109] and Imagic [110] provide important preparatory steps by aligning a real image with an editable manifold in the model before stable text or structural control can be achieved. With the rise of few-step diffusion and distilled models, the tension between inversion quality and inversion speed becomes more pronounced. Optimization-based inversion can provide accurate reconstructions but is computationally expensive, and it may create a stronger conflict between reconstruction fidelity and editability in few-step settings. ReNoise introduces iterative renoising during inversion to correct approximation error in the forward trajectory, improving inversion fidelity without substantially increasing the number of operations while maintaining editability across samplers and accelerated models [111]. Negative-prompt Inversion advances inversion from iterative optimization to a forward procedure by constructing negative prompts that yield approximately equivalent reconstructions, significantly reducing inversion time and improving usability for interactive editing [112]. Collectively, these results highlight an engineering consideration: when a collaborative control system treats real image editing as a standard workflow, inversion should be treated as a key control component that is plug-and-play, accelerable, and evaluable rather than an implicit pre-processing step. For non-rigid editing and multi-image consistency, MasaCtrl modifies self-attention into mutual self-attention. Without fine-tuning, it enables query-based reuse of appearance and texture from a source image and provides a mechanistic compromise between consistency and editability that combines naturally with structural control tools, such as ControlNet and T2I-Adapter [113]. Finally, Plug-and-Play Diffusion Features formulates image-to-image tasks through feature reuse. It unifies image editing as test-time feature guidance injection and supports structure guidance from sketches to real

images without training or fine-tuning, covering an important portion of the inference-time overlay spectrum for feature-level collaboration [114].

In personalization and multi-concept composition pathways, the objective is to address persistent difficulties of multi-condition collaboration through parameter efficiency and composability. Textual Inversion [115] learns new word embeddings from a small number of examples for concept injection. DreamBooth [116] fine-tunes the model for subject-driven generation but can entangle concepts with background attributes. Custom Diffusion [117] updates only a small subset of cross-attention parameters and supports joint training and closed-form merging of multiple concepts, improving scalability for multi-concept collaboration. SVDiff [118] fine-tunes only singular values to define a compact parameter space and uses data augmentation to mitigate overfitting and language drift, improving stability for multi-subject collaboration. Perfusion [119] uses key locking and gated rank-one updates to reduce personalization footprint and exposes explicit strength modulation under multi-concept composition. From a systems perspective, personalization modules inject persistent constraints for identity or instances into the backbone, whereas the core challenge in multi-concept composition is to prevent representational contamination and attention competition among concepts within shared cross-attention. SVDiff reduces drift risk by restricting the trainable subspace [118]. Perfusion anchors new concepts to hyper-classes through key locking and gating while providing an explicit interface for strength modulation [119]. These designs shift composability from heuristic tuning toward structured mechanisms. Concept Sliders further parameterizes controllability as continuous sliders along low-rank directions, enabling interpretable strength calibration with limited interference with other attributes. This provides a tunable and schedulable control primitive for collaborative systems and facilitates integration with multi-condition schedulers [84].

Cross-image consistency extends personalization from one output to a sequence. StoryDiffusion shares subject information through Consistent Self-Attention and uses a semantic-space motion predictor for transitions [120]. Unlike DreamBooth-style tuning, it does not store a subject by updating the backbone. Unlike independent image-prompt adapters, it couples tokens across generated frames. This design improves long-range identity and clothing consistency, but memory, error propagation, and temporal planning become more important as sequence length increases.

The broader progression is therefore hierarchical rather than chronological. Early single-branch methods established reliable spatial reachability but scaled poorly with the number of controls. Adapters reduced parameter cost, then decoupled attention channels reduced competition between image and text conditions. Multi-path methods improved regional independence at the expense of repeated inference. Dynamic controllers and auditor loops subsequently addressed conflicts that fixed fusion could not detect. No family dominates every criterion: branches favor strict geometry, adapters favor reuse, planners favor compositional semantics, and closed loops favor diagnosability. Their relative performance follows from the failure mode they are designed to expose and correct.

3.3. Zero-Shot/Few-Shot Adaptation for New Condition Extensions

The core goal of zero-shot or few-shot adaptation is to connect new conditions to existing generation pipelines in a composable manner without large-scale retraining of the backbone, thereby reducing the cost of condition extension and improving cross-task reuse. Among training-free or near-training-free inference-time approaches, SDEdit is an important baseline that leverages an SDE perspective to enable progressive editing while preserving the structure of an input image [104]. DiffEdit reduces annotation and interaction costs through an editing workflow driven by automatic mask inference [105].

pix2pix-zero improves edit directionality in a zero-shot setting by identifying directions in the text embedding space and reduces the need for prompt engineering [75]. MaskSketch injects sketch structure as sampling-time constraints, improving structural consistency while reducing reliance on paired data [107]. CycleNet introduces cycle-consistency regularization into diffusion-based image manipulation and uses self-consistency constraints to mitigate manipulation drift under unpaired data [88]. GDCC further strengthens stability for controllable manipulation through cycle consistency and contrastive constraints, improving reliability under inference-time and weakly supervised settings [89].

In this context, training-free does not imply an absence of learning signals. Instead, it shifts learning from parameter updates to inference trajectories. Typical approaches include applying differentiable constraints to attention or features and optimizing them iteratively during sampling, and using external models, such as detection, segmentation, VQA, or similarity models, to convert condition satisfaction into computable feedback that is injected into sampling or used for lightweight fine-tuning. As a result, zero-shot methods are often naturally coupled with evaluators or reward models and fit well into closed-loop scheduling within collaborative systems [75,105].

Zero-shot adaptation for identity and instance control has also emerged as plug-and-play condition modules and unified representation injection. InstantBooth learns image-to-token mappings together with a small set of adapter layers, enabling personalized generation without test-time fine-tuning while preserving text controllability [121]. InstantID injects identity semantics together with weak spatial conditions, such as facial landmarks, enabling identity-preserving generation from a single reference image and remaining compatible with multiple styles and downstream tasks [122]. PhotoMaker aggregates multiple identity references into a unified identity representation through stacked identity embeddings, improving identity fidelity while enhancing text controllability and inference efficiency. It also introduces identity-oriented data organization pipelines to support scalable training [123]. IP-Adapter maps image prompts to lightweight adapters, allowing image conditions to collaborate in parallel with text and structural control [96]. InstantID emphasizes explicit injection of identity semantics together with weak geometry, PhotoMaker emphasizes aggregation of multiple references into a unified identity representation for robustness, and IP-Adapter provides a more general image-prompt channel. Collectively, these methods indicate that identity-consistent generation benefits from moving identity constraints beyond pure text alignment into an independent and controllable representation pathway to reduce contamination and competition with region text and structural conditions in shared cross-attention.

Low-rank and parameter-efficient adaptation provides a stable engineering route for extending new conditions. LoRA is widely used as a general low-rank mechanism for lightweight customization of diffusion models. LoRA-Composer extends this idea to training-free multi-concept composition by injecting and isolating multiple pretrained concept LoRAs, reducing concept suppression and confusion [83]. SVDiff [118] and Perfusion [119] define smaller and more controllable update sets through singular-value subspaces and rank-one edits, respectively, improving multi-concept collaboration and reducing overfitting. DisenBooth introduces disentangled objectives for identity-related and identity-irrelevant embeddings to mitigate entanglement in subject-driven generation, allowing identity preservation and text controllability to be optimized jointly within a single system [82]. StyleDrop achieves single-image style personalization by fine-tuning a small subset of parameters and can improve style consistency through iterative feedback. The value of parameter-efficient methods in collaborative control systems lies in hot-swappability and versionability. When control modules must be stacked frequently in multi-condition settings or replaced frequently in multi-style and multi-brand scenarios,

small-footprint updates, such as LoRA, rank-one updates, and singular-value subspaces, are better suited to A/B testing, rollback, and multi-asset management in production [119].

For system-level scaling of collaboration, SDXL provides a stronger backbone that increases the capacity of multi-condition toolchains and has become a platform for adapters and personalization methods [45]. DynamicControl studies adaptive condition scheduling under multi-condition inputs from the perspective of condition selection and conflict management, indicating a transition from static composition to dynamic orchestration. ColLaFuse proposes a distributed client–server collaborative training framework that reduces on-device burdens while improving privacy properties, providing an implementation path for collaborative control systems in edge scenarios [124].

3.4. Computational Complexity and Scalability Considerations

Computational complexity is a critical but often under-discussed factor in collaborative control. When multiple control signals are introduced into a diffusion model, the total cost is no longer determined only by the pretrained denoising backbone. It also depends on auxiliary control branches, adapters, attention manipulation modules, region-wise denoising paths, dynamic selection policies, and external planning modules. Therefore, the computational cost of collaborative control can be approximately described as

$$C_{\text{total}} \approx C_{\text{base}} + \sum_{i=1}^N C_{\text{ctrl}}^{(i)} + C_{\text{fusion}} + C_{\text{sched}}, \quad (19)$$

where N is the number of active control components and i indexes one component. The term C_{base} is the cost of the base diffusion model, and $C_{\text{ctrl}}^{(i)}$ is the additional cost of component i . The terms C_{fusion} and C_{sched} denote fusion or arbitration and dynamic scheduling or external planning, respectively. More specifically, the source of control cost differs across technical pathways:

$$C_{\text{ctrl}}^{(i)} = \begin{cases} C_{\text{branch}}^{(i)} & \text{branch-based control,} \\ C_{\text{adapter}}^{(i)} & \text{adapter-based control,} \\ C_{\text{attn}}^{(i)} & \text{attention-based control,} \\ C_{\text{opt}}^{(i)} & \text{training-free optimization.} \end{cases} \quad (20)$$

This formulation highlights that scalability should be evaluated jointly with controllability and visual fidelity, especially when the number of control conditions increases.

As shown in Table 4, the main pathways have distinct cost structures. Branch-based methods provide strong structural control, but additional branches increase parameters and latency approximately linearly. Adapters reduce storage and reuse the frozen backbone, although their control can weaken under difficult combinations. MultiDiffusion avoids training but repeats denoising over regions or windows. UniCombine instead shares one FLUX trajectory and adds relatively small conditional modules. The source papers do not provide sufficiently complete inference-memory and wall-clock measurements under matched settings for a meaningful table-level comparison. Table 4 therefore retains source-supported parameter, training, and asymptotic scaling information, while resolution, sampler, denoising steps, precision, hardware, memory, and latency remain necessary items for future standardized reporting. For reproducibility, Table 5 complements this cost comparison by listing official public implementations of the representative control methods.

Table 4. Complexity and scalability of representative collaborative control pathways.

Method	Parameter Overhead	Train?	Reported Training Hardware/Time	Latency Complexity Scaling
ControlNet [6]	approximately one trainable encoder copy per control	Yes	1 × RTX 3090 Ti (24 GB), 5 days for the reported depth model	$O(KC_{\text{branch}})$ for K independently stacked branches
T2I-Adapter [7]	77 M; compressed variants 18 M/5 M	Yes	4 × Tesla V100 32 GB, within 3 days	one backbone pass plus $O(KC_{\text{adapter}})$
GLIGEN [68]	gated grounding layers; total count NR	Yes	NR	one trajectory; attention cost grows with grounding tokens
MultiDiffusion [76]	0 task-specific parameters	No	Not applicable	$O(WC_{\text{base}})$ for W overlapping denoising windows or paths
DynamicControl [70]	condition evaluator plus multi-control adapter	Yes	NR	evaluation over K inputs, then denoising with the selected subset
UniCombine [79]	29 M training-free/44 M training-based for two conditions	Optional	16 × V100, 30,000 steps for training-based mode	one shared trajectory; conditional-attention cost grows with condition tokens

Note: Train? indicates whether method-specific parameter training is required, and NR means not reported in the source paper. Training hardware and duration are descriptive rather than directly comparable because datasets, resolutions, precision, and optimization settings differ. C_{base} , C_{branch} , and C_{adapter} denote the costs of one base denoising pass, one control branch, and one adapter, respectively.

Table 5. Official public implementations for representative controllable and collaborative diffusion methods.

Method	Reproducibility Role	Official Repository
ControlNet [6]	branch-based structural-control baseline	https://github.com/lllyasviel/ControlNet (accessed on 21 July 2026)
T2I-Adapter [7]	lightweight composable-adapter baseline	https://github.com/TencentARC/T2I-Adapter (accessed on 21 July 2026)
GLIGEN [68]	grounded box and region control	https://github.com/gligen/GLIGEN (accessed on 21 July 2026)
MultiDiffusion [76]	multi-region and panorama inference	https://github.com/omerbt/MultiDiffusion (accessed on 21 July 2026)
ControlNet++ [87]	consistency-feedback control	https://github.com/liming-ai/ControlNet_Plus_Plus (accessed on 21 July 2026)
SLD [93]	LLM-audited closed-loop correction	https://github.com/tsunghan-wu/SLD (accessed on 21 July 2026)
RPG [94]	multimodal-LLM regional planning	https://github.com/YangLing0818/RPG-DiffusionMaster (accessed on 21 July 2026)
StoryDiffusion [120]	cross-image subject consistency	https://github.com/HVision-NKU/StoryDiffusion (accessed on 21 July 2026)
OmniGen [51]	unified generation and editing	https://github.com/VectorSpaceLab/OmniGen (accessed on 21 July 2026)
OminiControl [52]	parameter-efficient DiT condition injection	https://github.com/Yuanshi9815/OminiControl (accessed on 21 July 2026)
FLUX.1 [50]	rectified-flow Transformer backbone	https://github.com/black-forest-labs/flux (accessed on 21 July 2026)

Note: Repository ownership and accessibility were verified on 13 July 2026. The table identifies runnable starting points; exact environments, model licenses, and checkpoints must still be checked before reuse.

4. Application Scenarios of Collaborative Control Methods

From a task taxonomy perspective, controllable diffusion covers not only spatial structure control, including layout, pose, depth, and segmentation, and personalized generation, such as subject or style customization, but also extends to view-consistent generation with 3D awareness, context- or example-driven generation in the form of in-context editing, and practical directions such as text rendering and layout for real media production. Different tasks impose different requirements on structural consistency, cross-condition conflict management, and verifiable constraints. As a result, application-facing collaborative control is typically realized through systematic matching between control interfaces and task constraints, together with engineering trade-offs that prioritize the most critical failure modes.

4.1. Daily Creation and Commercial Content Production

In everyday image editing and commercial content pipelines, the value of collaborative control is primarily reflected in predictable local editability, cross-condition consistency, and high-throughput inference deployment. Inference-time editing paradigms, including SDEdit [104], DiffEdit [105], Prompt-to-Prompt [102], and InstructPix2Pix [103], are widely used for rapid modification, instruction-based retouching, and low-barrier interactive editing. Multi-region and layout control methods, exemplified by MultiDiffusion, are better suited to high-frequency production needs, such as multi-canvas e-commerce posters, partitioned content generation, and panoramic outpainting [76]. Interactive geometric editing methods, such as DragDiffusion, support drag-based structural adjustment and localized replacement, reducing the interaction burden for non-expert users [108]. In commercial pipelines, control methods are typically adopted as reusable workflow components rather than as one-off visual effects. For example, MultiDiffusion [76] supports reusable layout templates, Prompt-to-Prompt and Null-text inversion align well with lightweight editing instructions, such as batch recoloring or motif replacement, and DragDiffusion [108] is better suited to interactive refinement of key assets. Consequently, practical deployment of collaborative control usually follows a closed loop of automated generation and manual correction rather than fully replacing conventional design workflows.

From the perspective of coordinating structural conditions with generation quality, structured condition interfaces, such as ControlNet [6] and T2I-Adapter [7], simplify integration of edges, pose, depth, and segmentation into online toolchains, enabling predictable structural constraints together with stylized synthesis. In multi-subject and complex-prompt settings, attention manipulation and guidance-based methods, such as Attend-and-Excite [99] and SEGA [125], improve token-level alignment and provide more stable and tunable control for batch creation. Object-level transfer and composition methods, including AnyDoor [126] and Paint by Example [127], further support commercial scenarios, such as product replacement, prop composition, and brand style alignment, by providing reusable object-centric capabilities [64].

4.2. Precision Modeling and Optimization in Professional Domains

Professional domains emphasize structural correctness, cross-condition consistency, and satisfaction of task-specific constraints. In these settings, collaborative control is typically realized through combinations of structured conditions, domain priors, and generative backbones that enable verifiable and precise synthesis. For high-stakes scenarios with stringent reliability requirements, the central value of collaborative control is not only photorealism. Instead, domain constraints are injected into the diffusion process through structured conditions, external priors, or verifiable objective functions, ensuring that generated outputs satisfy interpretable, reproducible, and auditable criteria. These domains therefore rely more on verifiable constraints than on increasingly powerful backbones alone. Control signals often form a closed loop with downstream evaluators, such as detection, segmentation, or planning constraint checkers, and failure modes must remain diagnosable, whether they originate from condition conflicts, sampling drift, or domain shift. This helps explain why mechanisms such as cycle consistency, reward feedback, and task duality between generation and detection are particularly prevalent in professional applications.

In strongly constrained systems, including autonomous driving and unmanned aerial vehicles, diffusion models are increasingly used as generative planners or predictors that operate in coordination with rules, perception, and control modules. The progression is from sampling candidate trajectories to verifying and refining them against explicit objectives. Diffusion-ES combines diffusion-generated trajectory distributions with sampling-based search for test-time reward optimization, whereas CGD separates feasible path

generation from proxy-based time parameterization to enforce dynamics and collision constraints [128,129]. These designs show how generative diversity becomes useful only when paired with an external feasibility or reward mechanism.

In medical imaging and clinical assistance, the control interface becomes progressively more explicit. Med-DDPM uses semantic conditions to guide 3D MRI generation and support segmentation-oriented augmentation, while segmentation-guided diffusion uses anatomical masks to make the required structure directly verifiable [130,131]. This progression from semantic conditioning to explicit anatomical support reflects the stronger auditability required in clinical settings.

In remote sensing and geospatial intelligence, diffusion models are often used for cross-sensor conditional translation to support cloud removal, cross-modal completion, and disaster assessment. SAR-conditioned optical generation first established a mechanism for preserving cross-sensor structure while recovering optical texture; subsequent sampling acceleration and inference optimization then shifted attention from reconstruction quality toward operational throughput [132,133]. The methodological sequence therefore moves from cross-modal fidelity to deployable cross-sensor generation.

Digital humans, virtual avatars, and identity-consistent generation emphasize the joint satisfaction of identity preservation and editability. InstantID introduces a plug-and-play IdentityNet that injects identity semantics together with weak spatial conditions into pretrained diffusion models, enabling identity-preserving generation from a single facial reference [122]. PhotoMaker encodes multiple identity references into a unified identity representation via stacked identity embeddings, achieving high identity fidelity while improving text controllability and inference efficiency [123]. ConsistentID further targets fine-grained facial attribute prompts by strengthening disentanglement between identity consistency and editable attributes [134].

In architecture, interior design, and urban digital twin scenarios, collaborative control is primarily defined by translation of geometric and functional constraints into learnable or inferable structural conditions. LayoutDiffusion strengthens joint control over global layout and object-level detail, improving interpretability in complex multi-object scene generation [85].

In robotic manipulation and simulation environment construction, collaborative control with diffusion models often appears as integration of perceptual conditions, action distribution generation, and temporal closed-loop control. Diffusion Policy represents robot policies as conditional denoising diffusion processes, learning the score of action distributions and performing inference via Langevin dynamics to generate multimodal high-dimensional actions [135].

In fashion and virtual try-on, the methodological progression moves from coarse multimodal specification toward appearance fidelity, efficient synthesis, and editable production workflows. Multimodal Garment Designer establishes joint conditioning with text, pose, and garment sketches, defining the garment and body structure before synthesis [136]. IDM-VTON and GarDiff then strengthen preservation of garment-specific appearance by separating high-level semantics from low-level texture or by introducing garment-focused adapters and appearance objectives [137,138]. CAT-DM addresses a different bottleneck by combining ControlNet-style conditioning with GAN-based initialization to reduce the required sampling process [139]. Building on these foundations, SGDiff, structure-aware reference transfer, TexFit, and DiCTI extend control toward style guidance, reference-based design, text-driven editing, and inpainting-based variation [140–143]. Together, these methods form a layered pipeline in which pose and silhouette establish geometry, reference or adapter features preserve garment identity and texture, and editing interfaces support iterative design rather than one-shot generation.

4.3. Data Support and Evaluation for Model Innovation

A key role of collaborative control in research and engineering validation is to support controllable synthetic data pipelines, evaluation systems for controllability and consistency, and reproducible paradigms for innovation validation and ablation. In both research and engineering settings, collaborative control methods are often used as synthetic data engines and complex scenario constructors to compensate for scarce annotations, cover long-tail cases, and evaluate cross-condition generalization systematically. Wu et al. proposed DiffuMask, which leverages cross-attention signals in pretrained text–image diffusion models to generate pixel-level semantic masks alongside synthetic images, enabling low-cost construction of segmentation training data [144]. Chen et al. proposed DiffusionDet, which formulates object detection as a denoising diffusion process from noisy boxes to ground-truth boxes [145]. Saragih et al. constructed diffusion-generated annotated synthetic data pipelines for medical segmentation tasks [146]. Alemohammad et al. proposed SIMS, which uses self-generated data for self-improving training and provides a practical pathway for continual optimization of diffusion models using synthetic data [147]. Methodologically, the value of synthetic data is not limited to visual realism. Controllability and traceability are equally important. When control signals, such as boxes, segmentation, relations, and attributes, are recordable and replayable, performance changes in downstream tasks can be attributed to specific control modules or conflict-management strategies, enabling more credible ablation and comparison.

At the evaluation level, HEIM proposes a comprehensive benchmarking framework that covers alignment, quality, aesthetics, originality, reasoning, knowledge, bias, toxicity, fairness, robustness, multilinguality, and efficiency, providing a unified coordinate system for analyzing benefit–risk trade-offs of collaborative control methods [148]. TIFA converts text consistency evaluation into interpretable visual question answering by automatically generating question–answer pairs and using VQA models to judge correctness, enabling diagnostic assessment of challenges such as counting, relations, and attribute binding [149]. T2I-CompBench focuses on compositionality through attribute binding, object relations, and complex combinations and proposes evaluation metrics and optimization strategies tailored to compositional generation [150]. GenAI-Bench constructs compositional and reasoning-oriented benchmarks that are closer to practical prompting needs and emphasizes unified measurement of higher-order reasoning and compositional abilities [151]. Beyond traditional alignment and quality metrics, recent evaluation trends increasingly emphasize object-level diagnosability and mitigation of benchmark drift. GenEval introduces an object-centric evaluation pipeline based on detection and attribute verification and explicitly measures common failure modes in controllable generation, including object co-occurrence, position, counting, and color [152]. DSG, inspired by formal semantics, organizes question–answer evaluation as dependency graphs to reduce hallucinated questions and inconsistent answers, improving reliability and coverage for fine-grained assessment of complex relations and attribute binding [153]. As foundation model capabilities improve rapidly, static benchmarks face the risk of evaluation drift, where both models and test sets become outdated. GenEval 2 addresses benchmark drift by introducing stronger compositionality and finer-grained primitive coverage, making evaluation more sustainable over time [154]. In collaborative control research, evaluation frameworks often determine whether research questions can be advanced meaningfully. Only when evaluation decomposes failures into localized dimensions, such as missing objects, counting errors, relation mismatches, or attribute binding drift, can improvements in control modules and scheduling strategies avoid collapsing into subjective judgments of visual quality. Interpretable evaluations, such as TIFA, therefore provide an optimizable error decomposition for collaborative control [99].

Table 6 summarizes task-oriented metrics for evaluating precise collaborative control. General image fidelity metrics, such as FID and aesthetic score, are useful for measuring distribution-level quality but cannot determine whether a generated image satisfies structural, spatial, semantic, or identity constraints. Text-image metrics such as CLIP score and TIFA are effective for semantic alignment, but they provide limited diagnosis of layout drift, object omission, attribute-binding errors, or condition conflicts. Therefore, structure-guided generation should be evaluated with edge, pose, depth, or segmentation metrics; grounding-based generation should be evaluated with object detection or box-alignment metrics; and subject-driven generation should be evaluated with identity similarity metrics such as CLIP-I or DINO. For collaborative control systems, efficiency indicators, including additional parameters, GPU memory, inference latency, and the number of denoising runs, are also necessary because multi-condition control often introduces nontrivial system-level overhead.

Table 6. Task-oriented metrics and benchmarks for precise collaborative control.

Evaluation Target	Representative Metrics	Benchmark/Paper	Use in Collaborative Control
Holistic T2I capability	alignment, quality, aesthetics, reasoning, fairness, robustness, efficiency	HEIM [148]	broad system-level assessment across 12 aspects and 62 scenarios
Text-image faithfulness	VQA-based TIFA score; counting, relation, and attribute checks	TIFA [149]	diagnoses semantic adherence beyond CLIP-style global similarity
Object compositionality	co-occurrence, position, count, color accuracy	GenEval [152]; GenEval 2 [154]	measures common failures in multi-object and layout-sensitive prompts
Fine-grained semantics	dependency-structured question answering	DSG [153]	improves relation-centric diagnosis for complex scenes
Structural control	Canny F1, HED SSIM, OpenPose mAP, depth RMSE, mIoU	DynamicControl [70]	evaluates whether edge, pose, depth, or segmentation conditions are preserved
Grounding and layout	YOLO AP/AP50/AP75, GLIP score, FID	GLIGEN [68]	checks object-region alignment together with image quality
Subject consistency	CLIP-I, DINO; plus FID/SSIM/CLIP-T	UniCombine [79]	evaluates identity preservation under subject-spatial composition
Efficiency	GPU memory, extra parameters, inference cost	UniCombine/T2I-Adapter/ControlNet [6,7,79]	captures deployment overhead introduced by additional controls

Metric selection should follow the declared control contract rather than benchmark popularity. First, researchers should identify the primary constraint and its failure unit. Binary contour preservation calls for Canny F1 with a stated spatial tolerance, while soft edge similarity is better reflected by HED SSIM. Metric depth requires RMSE after specifying units and alignment, whereas relative geometry requires scale-invariant depth error. Pose control should use keypoint mAP, segmentation should use mIoU, and box grounding should use IoU or detection AP.

Second, pair the direct control metric with a non-collapse metric. FID or an aesthetic score can reveal degraded visual quality, while TIFA, DSG, or GenEval can reveal omitted objects, incorrect counts, and attribute-binding failures. Identity-conditioned tasks should add CLIP-I or DINO similarity, but these scores should not substitute for pose or layout accuracy.

Third, multi-condition studies should report a metric vector rather than an unexplained weighted average. The vector should include every condition-specific score, the worst-satisfied condition, and performance under deliberately conflicting inputs. Pareto curves over control strength and perceptual quality are more informative than one operating point. This protocol reveals whether an apparent gain comes from suppressing a difficult condition.

Finally, deployment studies should report parameters, peak GPU memory, wall-clock latency, resolution, precision, sampler, denoising steps, batch size, and hardware. A task-

specific metric is meaningful only when its evaluator is reliable in that domain. Medical, remote-sensing, and industrial applications should therefore validate automatic scores against expert or downstream-task assessments.

Table 7 summarizes representative benchmarks and evaluation dimensions relevant to controllable diffusion models.

Table 7. Representative benchmarks and evaluation dimensions for precise and controllable diffusion models.

Benchmark	Target	Granularity	Scope	Signals	Limitation	Control Analysis Use
HEIM [148]	holistic evaluation of text-to-image systems	system/multi-dimensional level	primarily T2I	alignment, image quality, aesthetics, originality, reasoning, knowledge, bias, toxicity, fairness, robustness, multilinguality, efficiency	not specifically designed for structured control or condition arbitration	useful for global system-level assessment, but less diagnostic for fine-grained control errors
TIFA [149]	text-image faithfulness	object/attribute/relation level	primarily T2I	question answering derived from prompts and answered from generated images	depends on the reliability of the underlying VQA pipeline	useful for checking whether generated outputs satisfy explicit semantic conditions
T2I-CompBench [150]	composition-focused T2I evaluation	composition level	primarily T2I	attribute binding, object relationships, and complex compositions	less comprehensive outside composition scenarios	useful for analyzing semantic conflicts and compositional failures under collaborative control
GenAI-Bench [151]	composition and reasoning evaluation	image/video/reasoning level	text-to-visual generation	human preference, composition prompts, and automatic signals	less focused on explicit structural control interfaces	useful for analyzing whether stronger planning or reasoning modules benefit controllable generation
GenEval [152]	object-focused evaluation of generated images	object/relation level	primarily T2I	object co-occurrence, position, count, and color	limited primitive coverage relative to broader real-world control requirements	useful for measuring common precise-generation failures such as counting and layout errors
DSG [153]	reliability of fine-grained text-image evaluation	relation/semantic-graph level	primarily T2I	dependency-structured question generation and answering	evaluation still depends on external models	useful for relation-centric diagnosis in complex multi-entity scenes

4.4. Application Potential and Future Directions

Future collaborative control will be shaped by real-time deployment, distributed and privacy-preserving execution, and governance. Consistency Models and Latent Consistency Models replace long sampling chains with one-step or few-step mappings while retaining editing ability [155,156]. Adversarial Diffusion Distillation further distills large teachers into one-to-four-step generators, improving fidelity in low-step regimes [157]. Few-step generation is not merely acceleration: it changes which control interfaces remain feasible. When denoising has only a few iterations, sampling-time optimization and iterative attention refinement have limited room, so stronger early injection, intermediate alignment, or distillation constraints become necessary.

Efficiency work on SDXL, MobileDiffusion, and TinyFusion shows that stronger backbones, mobile-aware design, and Transformer-layer pruning can support interactive and edge scenarios [45,158,159]. In privacy-sensitive settings, CollaFuse reduces client-side computation and data exposure through split-learning-style collaborative training and inference [124]. For authenticity and governance, Stable Signature, Tree-Ring Watermarking, and SynthID demonstrate provenance mechanisms that should be treated as first-class components in pipelines involving control branches, adapters, distilled models, and cloud-edge partitioning [160–162].

5. Conclusions

Diffusion models provide a flexible framework for controllable image generation because external constraints can act at several stages of denoising. This survey distinguished controllable diffusion from collaborative control and organized the latter by representation, injection, arbitration, scheduling, feedback, and evaluation. The central transition is from isolated condition response to coordinated satisfaction of several measurable constraints. Backbones, controllers, planners, and evaluators must therefore be analysed as one system rather than as independent modules.

The first open challenge is reliable conflict management. Current systems usually assume that conditions are correct and mutually compatible, yet real inputs may disagree in geometry, semantics, identity, or spatial scope. Future controllers should estimate condition uncertainty, detect infeasible combinations, and expose priorities before sampling. Timestep schedules also require stronger theoretical and empirical grounding. A policy should explain why a constraint acts at a particular noise level and how that choice changes the attainable trade-off among constraints.

The second challenge is maintaining control under few-step generation. Long sampling trajectories permit repeated attention refinement, latent optimization, and evaluator feedback. Distilled one-step or four-step models provide much less correction time. Future work should therefore move critical structure earlier, distil controller behaviour together with the generator, and develop inversion methods that preserve both reconstruction and editability. Efficient collaboration also requires shared condition interfaces whose parameter, memory, and latency costs grow sublinearly with the number of controls.

The third challenge is evaluation and reproducibility. No universal scalar score can represent structural fidelity, semantic correctness, identity, visual quality, and efficiency simultaneously. Benchmarks should report task-specific metric vectors, worst-condition performance, deliberate conflict tests, and Pareto trade-offs. Evaluator failures must also be audited because an LLM, detector, or segmenter can convert its own bias into an incorrect control update. Public code, exact hardware, sampler settings, peak memory, and latency should become standard components of comparative reporting.

The most promising direction is a typed and auditable control protocol that connects conditions, spatial support, confidence, priority, activation time, and verification. Such a protocol could allow U-Net, DiT, FLUX, or future backbones to share controllers without hiding conflicts inside architecture-specific features. Closed-loop systems should also support abstention when constraints are infeasible and preserve a trace of planner decisions and evaluator feedback. Progress toward precise generation will depend less on adding another isolated condition and more on making coordination measurable, efficient, and trustworthy.

Author Contributions: Conceptualization, J.Q.; methodology, Q.Z.; formal analysis, W.X. and Y.W.; investigation, J.Q., W.X. and Y.W.; data curation, Q.Z. and Y.W.; validation, X.C.; project administration, Y.W.; writing—original draft preparation, J.Q. and Y.W.; writing—review and editing, W.X., Q.Z. and X.C.; visualization, J.Q., X.C. and Y.W.; supervision, Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Beijing Natural Science Foundation, via grant number 4232017.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors acknowledge the support received from the Beijing Natural Science Foundation (Project No. 4232017).

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

Abbreviations

AP	Average Precision
CFG	Classifier-Free Guidance
CLIP	Contrastive Language–Image Pre-Training
DDIM	Denoising Diffusion Implicit Model
DDPM	Denoising Diffusion Probabilistic Model
DINO	Self-Distillation with No Labels
DiT	Diffusion Transformer
DSG	Davidsonian Scene Graph
FID	Fréchet Inception Distance
GPU	Graphics Processing Unit
HED	Holistically Nested Edge Detection
I2I	Image-to-Image
LDM	Latent Diffusion Model
LLM	Large Language Model
LoRA	Low-Rank Adaptation
mAP	Mean Average Precision
mIoU	Mean Intersection over Union
MMDiT	Multimodal Diffusion Transformer
MSE	Mean Squared Error
NR	Not Reported
ODE	Ordinary Differential Equation
RMSE	Root Mean Squared Error
RPG	Recaption, Plan, and Generate
SAR	Synthetic Aperture Radar
SDE	Stochastic Differential Equation
SDXL	Stable Diffusion XL
SLD	Self-Correcting LLM-Controlled Diffusion Models
SSIM	Structural Similarity Index Measure
T2I	Text-to-Image
TIFA	Text-to-Image Faithfulness
VAE	Variational Autoencoder
VQA	Visual Question Answering

References

1. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
2. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
3. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv* **2020**, arXiv:2011.13456. [[CrossRef](#)]
4. Zhang, N.; Tang, H. Text-to-image synthesis: A decade survey. *arXiv* **2024**, arXiv:2411.16164. [[CrossRef](#)]
5. Cao, P.; Zhou, F.; Song, Q.; Yang, L. Controllable Generation with Text-to-Image Diffusion Models: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, *48*, 4771–4791. [[CrossRef](#)] [[PubMed](#)]
6. Zhang, L.; Rao, A.; Agrawala, M. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 3836–3847. [[CrossRef](#)]
7. Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; Shan, Y. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 4296–4304. [[CrossRef](#)]
8. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114. [[CrossRef](#)]

9. LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; Huang, F.J. Energy-Based Models. In *Predicting Structured Data*; Bakir, G., Hofmann, T., Scholkopf, B., Smola, A.J., Taskar, B., Vishwanathan, S.V.N., Eds.; MIT Press: Cambridge, MA, USA, 2007; pp. 191–246. [\[CrossRef\]](#)
10. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [\[CrossRef\]](#)
11. Dinh, L.; Sohl-Dickstein, J.; Bengio, S. Density estimation using real nvp. *arXiv* **2016**, arXiv:1605.08803. [\[CrossRef\]](#)
12. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *Proc. Mach. Learn. Res.* **2015**, *37*, 2256–2265.
13. Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; Zhu, J. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 5775–5787. [\[CrossRef\]](#)
14. Karras, T.; Aittala, M.; Aila, T.; Laine, S. Elucidating the design space of diffusion-based generative models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 26565–26577. [\[CrossRef\]](#)
15. Lipman, Y.; Chen, R.T.; Ben-Hamu, H.; Nickel, M.; Le, M. Flow matching for generative modeling. *arXiv* **2022**, arXiv:2210.02747. [\[CrossRef\]](#)
16. Liu, X.; Gong, C.; Liu, Q. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. *arXiv* **2022**, arXiv:2209.03003. [\[CrossRef\]](#)
17. Song, Y.; Ermon, S. Generative modeling by estimating gradients of the data distribution. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 11918–11930.
18. Jolicoeur-Martineau, A.; Piché-Taillefer, R.; Combes, R.T.d.; Mitliagkas, I. Adversarial score matching and improved sampling for image generation. *arXiv* **2020**, arXiv:2009.05475. [\[CrossRef\]](#)
19. De Bortoli, V.; Thornton, J.; Heng, J.; Doucet, A. Diffusion schrödinger bridge with applications to score-based generative modeling. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17695–17709. [\[CrossRef\]](#)
20. Song, J.; Meng, C.; Ermon, S. Denoising diffusion implicit models. *arXiv* **2020**, arXiv:2010.02502. [\[CrossRef\]](#)
21. Nichol, A.Q.; Dhariwal, P. Improved denoising diffusion probabilistic models. *Proc. Mach. Learn. Res.* **2021**, *139*, 8162–8171.
22. Salimans, T.; Ho, J. Progressive distillation for fast sampling of diffusion models. *arXiv* **2022**, arXiv:2202.00512. [\[CrossRef\]](#)
23. Zhao, W.; Bai, L.; Rao, Y.; Zhou, J.; Lu, J. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 49842–49869. [\[CrossRef\]](#)
24. Xue, S.; Liu, Z.; Chen, F.; Zhang, S.; Hu, T.; Xie, E.; Li, Z. Accelerating diffusion sampling with optimized time steps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 8292–8301. [\[CrossRef\]](#)
25. Liu, X.; Park, D.H.; Azadi, S.; Zhang, G.; Chopikyan, A.; Hu, Y.; Shi, H.; Rohrbach, A.; Darrell, T. More control for free! image synthesis with semantic diffusion guidance. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 289–299. [\[CrossRef\]](#)
26. Ho, J.; Salimans, T. Classifier-free diffusion guidance. *arXiv* **2022**, arXiv:2207.12598. [\[CrossRef\]](#)
27. Shen, D.; Song, G.; Xue, Z.; Wang, F.Y.; Liu, Y. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 9370–9379. [\[CrossRef\]](#)
28. 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, New Orleans, LA, USA, 21–24 June 2022; pp. 10684–10695. [\[CrossRef\]](#)
29. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: New York, NY, USA, 2015; pp. 234–241. [\[CrossRef\]](#)
30. Peebles, W.; Xie, S. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 4195–4205. [\[CrossRef\]](#)
31. Bao, F.; Nie, S.; Xue, K.; Cao, Y.; Li, C.; Su, H.; Zhu, J. All are worth words: A vit backbone for diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 22669–22679. [\[CrossRef\]](#)
32. Zheng, H.; Nie, W.; Vahdat, A.; Anandkumar, A. Fast training of diffusion models with masked transformers. *arXiv* **2023**, arXiv:2306.09305. [\[CrossRef\]](#)
33. Gao, S.; Zhou, P.; Cheng, M.M.; Yan, S. Masked Diffusion Transformer Is a Strong Image Synthesizer. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; IEEE: New York, NY, USA, 2023; pp. 23107–23116. [\[CrossRef\]](#)
34. Lu, Z.; Wang, Z.; Huang, D.; Wu, C.; Liu, X.; Ouyang, W.; Bai, L. Fit: Flexible vision transformer for diffusion model. *arXiv* **2024**, arXiv:2402.12376. [\[CrossRef\]](#)

35. Shen, T.; Yu, J.; Zhou, D.; Li, D.; Barsoum, E. E-MMDiT: Revisiting Multimodal Diffusion Transformer Design for Fast Image Synthesis under Limited Resources. *arXiv* **2025**, arXiv:2510.27135. [[CrossRef](#)]
36. Yang, L.; Huang, Z.; Zhang, Z.; Liu, Z.; Hong, S.; Zhang, W.; Yang, W.; Cui, B.; Zhang, L. Graphusion: Latent diffusion for graph generation. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 6358–6369. [[CrossRef](#)]
37. Jia, W.; Huang, M.; Chen, N.; Zhang, L.; Mao, Z. D²iT: Dynamic Diffusion Transformer for Accurate Image Generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 13–15 June 2025; pp. 12860–12870. [[CrossRef](#)]
38. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. *Proc. Mach. Learn. Res.* **2021**, *139*, 8748–8763.
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
40. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. *Proc. Mach. Learn. Res.* **2021**, *139*, 8821–8831.
41. Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Zhang, Q.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv* **2022**, arXiv:2211.01324. [[CrossRef](#)]
42. Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv* **2021**, arXiv:2112.10741. [[CrossRef](#)]
43. 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.* **2022**, *35*, 36479–36494. [[CrossRef](#)]
44. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67. [[CrossRef](#)]
45. 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. [[CrossRef](#)]
46. Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. Improving Image Generation with Better Captions. OpenAI Technical Report. 2023. Available online: <https://cdn.openai.com/papers/dall-e-3.pdf> (accessed on 21 July 2026).
47. Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv* **2023**, arXiv:2310.00426. [[CrossRef](#)]
48. Chen, J.; Ge, C.; Xie, E.; Wu, Y.; Yao, L.; Ren, X.; Wang, Z.; Luo, P.; Lu, H.; Li, Z. Pixart- σ : Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: New York, NY, USA, 2024; pp. 74–91. [[CrossRef](#)]
49. Li, Z.; Zhang, J.; Lin, Q.; Xiong, J.; Long, Y.; Deng, X.; Zhang, Y.; Liu, X.; Huang, M.; Xiao, Z.; et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv* **2024**, arXiv:2405.08748. [[CrossRef](#)]
50. Black Forest Labs. FLUX.1: Official Inference Code and Model Releases. GitHub Repository. 2024. Available online: <https://github.com/black-forest-labs/flux> (accessed on 13 July 2026).
51. Xiao, S.; Wang, Y.; Zhou, J.; Yuan, H.; Xing, X.; Yan, R.; Li, C.; Wang, S.; Huang, T.; Liu, Z. OmniGen: Unified Image Generation. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; IEEE: New York, NY, USA, 2025; pp. 13294–13304. [[CrossRef](#)]
52. Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; Wang, X. OminiControl: Minimal and Universal Control for Diffusion Transformer. In Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–25 October 2025; IEEE: New York, NY, USA, 2025; pp. 14940–14950. [[CrossRef](#)]
53. Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; Yoon, S. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 14347–14356. [[CrossRef](#)]
54. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134. [[CrossRef](#)]
55. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232. [[CrossRef](#)]
56. Huang, X.; Liu, M.Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 179–196. [[CrossRef](#)] [[PubMed](#)]
57. Lee, H.Y.; Tseng, H.Y.; Huang, J.B.; Singh, M.; Yang, M.H. Diverse image-to-image translation via disentangled representations. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 36–52. [[CrossRef](#)]

58. Kim, G.; Kwon, T.; Ye, J.C. Diffusionclip: Text-guided diffusion models for robust image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2426–2435. [\[CrossRef\]](#)
59. Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; Norouzi, M. Palette: Image-to-image diffusion models. In Proceedings of the ACM SIGGRAPH 2022 Conference Proceedings, Vancouver, BC, Canada, 7–11 August 2022; pp. 1–10. [\[CrossRef\]](#)
60. Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D.J.; Norouzi, M. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4713–4726. [\[CrossRef\]](#) [\[PubMed\]](#)
61. Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; Van Gool, L. Diffir: Efficient diffusion model for image restoration. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 13095–13105. [\[CrossRef\]](#)
62. Yue, Z.; Wang, J.; Loy, C.C. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 13294–13307. [\[CrossRef\]](#)
63. Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11461–11471. [\[CrossRef\]](#)
64. Avrahami, O.; Lischinski, D.; Fried, O. Blended diffusion for text-driven editing of natural images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18208–18218. [\[CrossRef\]](#)
65. Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; Courville, A. Film: Visual reasoning with a general conditioning layer. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, 3942–3951. [\[CrossRef\]](#)
66. Yang, L.; Huang, Z.; Song, Y.; Hong, S.; Li, G.; Zhang, W.; Cui, B.; Ghanem, B.; Yang, M.H. Diffusion-based scene graph to image generation with masked contrastive pre-training. *arXiv* **2022**, arXiv:2211.11138. [\[CrossRef\]](#)
67. Voynov, A.; Aberman, K.; Cohen-Or, D. Sketch-guided text-to-image diffusion models. In Proceedings of the ACM SIGGRAPH 2023 Conference Proceedings, Los Angeles, CA, USA, 23 July 2023; pp. 1–11. [\[CrossRef\]](#)
68. Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; Lee, Y.J. Gligen: Open-set grounded text-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 22511–22521. [\[CrossRef\]](#)
69. Bhat, S.F.; Mitra, N.; Wonka, P. Loosecontrol: Lifting controlnet for generalized depth conditioning. In Proceedings of the ACM SIGGRAPH 2024 Conference Papers, Denver, CO, USA, 28 July–1 August 2024; pp. 1–11. [\[CrossRef\]](#)
70. He, Q.; Peng, J.; Xu, P.; Jiang, B.; Hu, X.; Luo, D.; Liu, Y.; Wang, Y.; Wang, C.; Li, X.; et al. DynamicControl: Adaptive Condition Selection for Improved Text-to-Image Generation. *arXiv* **2024**, arXiv:2412.03255. [\[CrossRef\]](#)
71. Zhao, S.; Chen, D.; Chen, Y.C.; Bao, J.; Hao, S.; Yuan, L.; Wong, K.Y.K. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 11127–11150. [\[CrossRef\]](#)
72. Gafni, O.; Polyak, A.; Ashual, O.; Sheynin, S.; Parikh, D.; Taigman, Y. Make-a-scene: Scene-based text-to-image generation with human priors. In *Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Springer: New York, NY, USA, 2022; pp. 89–106. [\[CrossRef\]](#)
73. Avrahami, O.; Hayes, T.; Gafni, O.; Gupta, S.; Taigman, Y.; Parikh, D.; Lischinski, D.; Fried, O.; Yin, X. Spatext: Spatio-textual representation for controllable image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 18370–18380. [\[CrossRef\]](#)
74. Liu, N.; Li, S.; Du, Y.; Torralba, A.; Tenenbaum, J.B. Compositional visual generation with composable diffusion models. In *Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Springer: New York, NY, USA, 2022; pp. 423–439. [\[CrossRef\]](#)
75. Parmar, G.; Kumar Singh, K.; Zhang, R.; Li, Y.; Lu, J.; Zhu, J.Y. Zero-shot image-to-image translation. In Proceedings of the ACM SIGGRAPH 2023 Conference Proceedings, Los Angeles, CA, USA, 6–10 August 2023; pp. 1–11. [\[CrossRef\]](#)
76. Bar-Tal, O.; Yariv, L.; Lipman, Y.; Dekel, T. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. *arXiv* **2023**, arXiv:2302.08113. [\[CrossRef\]](#)
77. Huang, L.; Chen, D.; Liu, Y.; Shen, Y.; Zhao, D.; Zhou, J. Composer: Creative and controllable image synthesis with composable conditions. *arXiv* **2023**, arXiv:2302.09778. [\[CrossRef\]](#)
78. Hu, M.; Zheng, J.; Liu, D.; Zheng, C.; Wang, C.; Tao, D.; Cham, T.J. Cocktail: Mixing multi-modality control for text-conditional image generation. In Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; pp. 32424–32444. [\[CrossRef\]](#)
79. Wang, H.; Peng, J.; He, Q.; Yang, H.; Jin, Y.; Wu, J.; Hu, X.; Pan, Y.; Gan, Z.; Chi, M.; et al. UniCombine: Unified Multi-Conditional Combination with Diffusion Transformer. In *Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025*; IEEE: New York, NY, USA, 2025; pp. 18325–18334. [\[CrossRef\]](#)

80. Yang, H.; Han, W.; Zhou, Y.; Shen, J. DC-ControlNet: Decoupling Inter- and Intra-Element Conditions in Image Generation with Diffusion Models. In *Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–20 October 2025*; IEEE: New York, NY, USA, 2025; pp. 1–10. [\[CrossRef\]](#)
81. Xie, Y.; Feng, F.; Shi, R.; Wang, J.; Rui, Y.; Geng, X. DivControl: Knowledge Diversion for Controllable Image Generation. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 27108–27116. [\[CrossRef\]](#)
82. Chen, H.; Zhang, Y.; Wu, S.; Wang, X.; Duan, X.; Zhou, Y.; Zhu, W. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. *arXiv* **2023**, arXiv:2305.03374. [\[CrossRef\]](#)
83. Yang, Y.; Wang, W.; Peng, L.; Song, C.; Chen, Y.; Li, H.; Yang, X.; Lu, Q.; Cai, D.; He, X.; et al. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *IEEE Trans. Image Process.* **2025**, *34*, 8145–8158. [\[CrossRef\]](#) [\[PubMed\]](#)
84. Gandikota, R.; Materzyńska, J.; Zhou, T.; Torralba, A.; Bau, D. Concept sliders: Lora adaptors for precise control in diffusion models. In *Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024*; Springer: New York, NY, USA, 2024; pp. 172–188. [\[CrossRef\]](#)
85. Zheng, G.; Zhou, X.; Li, X.; Qi, Z.; Shan, Y.; Li, X. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023*; pp. 22490–22499. [\[CrossRef\]](#)
86. Chen, H.; Gao, Y.; Zhou, M.; Wang, P.; Li, X.; Ge, T.; Zheng, B. Enhancing prompt following with visual control through training-free mask-guided diffusion. *arXiv* **2024**, arXiv:2404.14768. [\[CrossRef\]](#)
87. Li, M.; Yang, T.; Kuang, H.; Wu, J.; Wang, Z.; Xiao, X.; Chen, C. ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback. In *Proceedings of the Computer Vision—ECCV 2024, Milan, Italy, 29 September–4 October 2024*; Springer Nature: Cham, Switzerland, 2024; pp. 129–147. [\[CrossRef\]](#)
88. Xu, S.; Ma, Z.; Huang, Y.; Lee, H.; Chai, J. Cyclenet: Rethinking cycle consistency in text-guided diffusion for image manipulation. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 10359–10384. [\[CrossRef\]](#)
89. Cai, X.; Lai, Q.; Pei, G.; Shu, X.; Yao, Y.; Wang, W. Cycle-consistent learning for joint layout-to-image generation and object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 21–23 October 2025*; pp. 6797–6807. [\[CrossRef\]](#)
90. Li, J.; Li, B.; Tu, Z.; Liu, X.; Guo, Q.; Juefei-Xu, F.; Xu, R.; Yu, H. Light the night: A multi-condition diffusion framework for unpaired low-light enhancement in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024*; pp. 15205–15215. [\[CrossRef\]](#)
91. Yang, J.; Li, A.; Liao, X.; Masouros, C. Speeding-up symbol-level precoding using separable and dual optimizations. *IEEE Trans. Commun.* **2023**, *71*, 7056–7071. [\[CrossRef\]](#)
92. Hu, X.; Wang, R.; Fang, Y.; Fu, B.; Cheng, P.; Yu, G. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv* **2024**, arXiv:2403.05135. [\[CrossRef\]](#)
93. Wu, T.H.; Lian, L.; Gonzalez, J.E.; Li, B.; Darrell, T. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024*; pp. 6327–6336. [\[CrossRef\]](#)
94. Yang, L.; Yu, Z.; Meng, C.; Xu, M.; Ermon, S.; Cui, B. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal LLMs. *Proc. Mach. Learn. Res.* **2024**, *235*, 56704–56721.
95. Qin, C.; Zhang, S.; Yu, N.; Feng, Y.; Yang, X.; Zhou, Y.; Wang, H.; Niebles, J.C.; Xiong, C.; Savarese, S.; et al. UniControl: A Unified Diffusion Model for Controllable Visual Generation in the Wild. In *Proceedings of the Advances in Neural Information Processing Systems 36, New Orleans, LA, USA, 10–16 December 2023*; pp. 42961–42992. [\[CrossRef\]](#)
96. 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. [\[CrossRef\]](#)
97. He, X.; Zheng, J.; Fang, J.Z.; Piramuthu, R.; Bansal, M.; Ordonez, V.; Sigurdsson, G.A.; Peng, N.; Wang, X.E. FlexEControl: Flexible and efficient multimodal control for text-to-image generation. *arXiv* **2024**, arXiv:2405.04834. [\[CrossRef\]](#)
98. Xie, J.; Li, Y.; Huang, Y.; Liu, H.; Zhang, W.; Zheng, Y.; Shou, M.Z. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023*; pp. 7452–7461. [\[CrossRef\]](#)
99. Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; Cohen-Or, D. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph. (TOG)* **2023**, *42*, 148. [\[CrossRef\]](#)
100. Chen, J.; Huang, Y.; Lv, T.; Cui, L.; Chen, Q.; Wei, F. Textdiffuser: Diffusion models as text painters. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 9353–9387. [\[CrossRef\]](#)
101. Chen, J.; Huang, Y.; Lv, T.; Cui, L.; Chen, Q.; Wei, F. Textdiffuser-2: Unleashing the power of language models for text rendering. In *Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024*; Springer: New York, NY, USA, 2024; pp. 386–402. [\[CrossRef\]](#)

102. Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv* **2022**, arXiv:2208.01626. [\[CrossRef\]](#)
103. Brooks, T.; Holynski, A.; Efros, A.A. Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–26 June 2023; pp. 18392–18402. [\[CrossRef\]](#)
104. Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.Y.; Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv* **2021**, arXiv:2108.01073. [\[CrossRef\]](#)
105. Couairon, G.; Verbeek, J.; Schwenk, H.; Cord, M. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv* **2022**, arXiv:2210.11427. [\[CrossRef\]](#)
106. Huang, N.; Tang, F.; Dong, W.; Lee, T.Y.; Xu, C. Region-aware diffusion for zero-shot text-driven image editing. *arXiv* **2023**, arXiv:2302.11797. [\[CrossRef\]](#)
107. Bashkirova, D.; Lezama, J.; Sohn, K.; Saenko, K.; Essa, I. Masksketch: Unpaired structure-guided masked image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1879–1889. [\[CrossRef\]](#)
108. Shi, Y.; Xue, C.; Liew, J.H.; Pan, J.; Yan, H.; Zhang, W.; Tan, V.Y.; Bai, S. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 8839–8849. [\[CrossRef\]](#)
109. Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; Cohen-Or, D. Null-text inversion for editing real images using guided diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 6038–6047. [\[CrossRef\]](#)
110. Kavar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; Irani, M. Imagic: Text-based real image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 6007–6017. [\[CrossRef\]](#)
111. Garibi, D.; Patashnik, O.; Voynov, A.; Averbuch-Elor, H.; Cohen-Or, D. Renoise: Real image inversion through iterative noising. In *Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024*; Springer: New York, NY, USA, 2024; pp. 395–413. [\[CrossRef\]](#)
112. Miyake, D.; Iohara, A.; Saito, Y.; Tanaka, T. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. In *Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 28 February–4 March 2025*; IEEE: New York, NY, USA, 2025; pp. 2063–2072. [\[CrossRef\]](#)
113. Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; Zheng, Y. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 22560–22570. [\[CrossRef\]](#)
114. Tumanyan, N.; Geyer, M.; Bagon, S.; Dekel, T. Plug-and-play diffusion features for text-driven image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 1921–1930. [\[CrossRef\]](#)
115. Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A.H.; Chechik, G.; Cohen-Or, D. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. *arXiv* **2022**, arXiv:2208.01618. [\[CrossRef\]](#)
116. 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, Vancouver, BC, Canada, 18–22 June 2023; pp. 22500–22510. [\[CrossRef\]](#)
117. Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; Zhu, J.Y. Multi-Concept Customization of Text-to-Image Diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 1931–1941. [\[CrossRef\]](#)
118. Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; Yang, F. Svdif: Compact parameter space for diffusion fine-tuning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 7323–7334. [\[CrossRef\]](#)
119. Tewel, Y.; Gal, R.; Chechik, G.; Atzmon, Y. Key-locked rank one editing for text-to-image personalization. In Proceedings of the ACM SIGGRAPH 2023 Conference Proceedings, Los Angeles, CA, USA, 23 July 2023; pp. 1–11. [\[CrossRef\]](#)
120. Zhou, Y.; Zhou, D.; Cheng, M.M.; Feng, J.; Hou, Q. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation. In Proceedings of the Advances in Neural Information Processing Systems 37, Vancouver, BC, Canada, 10–15 December 2024; pp. 110315–110340. [\[CrossRef\]](#)
121. Shi, J.; Xiong, W.; Lin, Z.; Jung, H.J. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024*; IEEE: New York, NY, USA, 2024; pp. 8543–8552. [\[CrossRef\]](#)

122. Wang, Q.; Bai, X.; Wang, H.; Qin, Z.; Chen, A.; Li, H.; Tang, X.; Hu, Y. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv* **2024**, arXiv:2401.07519. [[CrossRef](#)]
123. Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.M.; Shan, Y. Photomaker: Customizing realistic human photos via stacked id embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 19–21 June 2024; pp. 8640–8650. [[CrossRef](#)]
124. Allmendinger, S.; Zipperling, D.; Struppek, L.; Kühl, N. Collafuse: Collaborative diffusion models. *arXiv* **2024**, arXiv:2406.14429. [[CrossRef](#)]
125. Brack, M.; Friedrich, F.; Hintersdorf, D.; Struppek, L.; Schramowski, P.; Kersting, K. Sega: Instructing text-to-image models using semantic guidance. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 25365–25389. [[CrossRef](#)]
126. Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; Zhao, H. Anydoor: Zero-shot object-level image customization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 6593–6602. [[CrossRef](#)]
127. Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; Wen, F. Paint by example: Exemplar-based image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 18381–18391. [[CrossRef](#)]
128. Yang, B.; Su, H.; Gkanatsios, N.; Ke, T.W.; Jain, A.; Schneider, J.; Fragkiadaki, K. Diffusion-ES: Gradient-free planning with diffusion for autonomous and instruction-guided driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 15342–15353. [[CrossRef](#)]
129. Kondo, K.; Tagliabue, A.; Cai, X.; Tewari, C.; Garcia, O.; Espitia-Alvarez, M.; How, J.P. Cgd: Constraint-guided diffusion policies for uav trajectory planning. *arXiv* **2024**, arXiv:2405.01758. [[CrossRef](#)]
130. Dorjsembe, Z.; Pao, H.K.; Odonchimed, S.; Xiao, F. Conditional diffusion models for semantic 3D brain MRI synthesis. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 4084–4093. [[CrossRef](#)] [[PubMed](#)]
131. Konz, N.; Chen, Y.; Dong, H.; Mazurowski, M.A. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Marrakesh, Morocco, 6–10 October 2024*; Springer: New York, NY, USA, 2024; pp. 88–98. [[CrossRef](#)]
132. Bai, X.; Pu, X.; Xu, F. Conditional diffusion for SAR to optical image translation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 4000605. [[CrossRef](#)]
133. Bai, X.; Xu, F. Accelerating diffusion for sar-to-optical image translation via adversarial consistency distillation. *arXiv* **2024**, arXiv:2407.06095. [[CrossRef](#)]
134. Huang, J.; Dong, X.; Song, W.; Chong, Z.; Tang, Z.; Zhou, J.; Cheng, Y.; Chen, L.; Li, H.; Yan, Y.; et al. Consistentid: Portrait generation with multimodal fine-grained identity preserving. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, *48*, 5639–5654. [[CrossRef](#)] [[PubMed](#)]
135. Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; Song, S. Diffusion policy: Visuomotor policy learning via action diffusion. *Int. J. Robot. Res.* **2025**, *44*, 1684–1704. [[CrossRef](#)]
136. Baldrati, A.; Morelli, D.; Cartella, G.; Cornia, M.; Bertini, M.; Cucchiara, R. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 23393–23402. [[CrossRef](#)]
137. Choi, Y.; Kwak, S.; Lee, K.; Choi, H.; Shin, J. Improving diffusion models for authentic virtual try-on in the wild. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 206–235. [[CrossRef](#)]
138. Wan, S.; Li, Y.; Chen, J.; Pan, Y.; Yao, T.; Cao, Y.; Mei, T. Improving virtual try-on with garment-focused diffusion models. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 184–199. [[CrossRef](#)] [[PubMed](#)]
139. Zeng, J.; Song, D.; Nie, W.; Tian, H.; Wang, T.; Liu, A.A. Cat-dm: Controllable accelerated virtual try-on with diffusion model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 8372–8382. [[CrossRef](#)]
140. Sun, Z.; Zhou, Y.; He, H.; Mok, P. Sgdif: A style guided diffusion model for fashion synthesis. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 8433–8442. [[CrossRef](#)]
141. Cao, S.; Chai, W.; Hao, S.; Zhang, Y.; Chen, H.; Wang, G. Diffashion: Reference-based fashion design with structure-aware transfer by diffusion models. *IEEE Trans. Multimed.* **2024**, *26*, 3962–3975. [[CrossRef](#)]
142. Wang, T.; Ye, M. Textfit: Text-driven fashion image editing with diffusion models. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 10198–10206. [[CrossRef](#)]
143. Lampe, A.; Stopar, J.; Jain, D.K.; Omachi, S.; Peer, P.; Štruc, V. Dicti: Diffusion-based clothing designer via text-guided input. In *Proceedings of the 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), Istanbul, Turkey, 27–31 May 2024*; IEEE: New York, NY, USA, 2024; pp. 1–9. [[CrossRef](#)]

144. Wu, W.; Zhao, Y.; Shou, M.Z.; Zhou, H.; Shen, C. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 1206–1217. [\[CrossRef\]](#)
145. Chen, S.; Sun, P.; Song, Y.; Luo, P. Diffusiondet: Diffusion model for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 19830–19843. [\[CrossRef\]](#)
146. Saragih, D.G.; Hibi, A.; Tyrrell, P.N. Using diffusion models to generate synthetic labeled data for medical image segmentation. *Int. J. Comput. Assist. Radiol. Surg.* **2024**, *19*, 1615–1625. [\[CrossRef\]](#) [\[PubMed\]](#)
147. Alemohammad, S.; Humayun, A.I.; Agarwal, S.; Collomosse, J.; Baraniuk, R. Self-improving diffusion models with synthetic data. *arXiv* **2024**, arXiv:2408.16333. [\[CrossRef\]](#)
148. Lee, T.; Yasunaga, M.; Meng, C.; Mai, Y.; Park, J.S.; Gupta, A.; Zhang, Y.; Narayanan, D.; Teufel, H.; Bellagente, M.; et al. Holistic evaluation of text-to-image models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 69981–70011. [\[CrossRef\]](#)
149. 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, Paris, France, 1–6 October 2023; pp. 20406–20417. [\[CrossRef\]](#)
150. Sun, K.; Huang, K.; Liu, X.; Wu, Y.; Xu, Z.; Li, Z.; Liu, X. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 8406–8416. [\[CrossRef\]](#)
151. Li, B.; Lin, Z.; Pathak, D.; Li, J.; Fei, Y.; Wu, K.; Ling, T.; Xia, X.; Zhang, P.; Neubig, G.; et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv* **2024**, arXiv:2406.13743. [\[CrossRef\]](#)
152. Ghosh, D.; Hajishirzi, H.; Schmidt, L. Geneval: An object-focused framework for evaluating text-to-image alignment. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 52132–52152. [\[CrossRef\]](#)
153. Cho, J.; Hu, Y.; Garg, R.; Anderson, P.; Krishna, R.; Baldridge, J.; Bansal, M.; Pont-Tuset, J.; Wang, S. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *arXiv* **2023**, arXiv:2310.18235. [\[CrossRef\]](#)
154. Kamath, A.; Chang, K.W.; Krishna, R.; Zettlemoyer, L.; Hu, Y.; Ghazvininejad, M. GenEval 2: Addressing Benchmark Drift in Text-to-Image Evaluation. *arXiv* **2025**, arXiv:2512.16853. [\[CrossRef\]](#)
155. Song, Y.; Dhariwal, P.; Chen, M.; Sutskever, I. Consistency Models. *Proc. Mach. Learn. Res.* **2023**, *202*, 32211–32252.
156. Luo, S.; Tan, Y.; Huang, L.; Li, J.; Zhao, H. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv* **2023**, arXiv:2310.04378. [\[CrossRef\]](#)
157. Sauer, A.; Lorenz, D.; Blattmann, A.; Rombach, R. Adversarial diffusion distillation. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 87–103. [\[CrossRef\]](#)
158. Zhao, Y.; Xu, Y.; Xiao, Z.; Jia, H.; Hou, T. Mobilediffusion: Instant text-to-image generation on mobile devices. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 225–242. [\[CrossRef\]](#)
159. Fang, G.; Li, K.; Ma, X.; Wang, X. Tinyfusion: Diffusion transformers learned shallow. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 18144–18154. [\[CrossRef\]](#)
160. Fernandez, P.; Couairon, G.; Jégou, H.; Douze, M.; Furon, T. The stable signature: Rooting watermarks in latent diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 13–15 June 2023; pp. 22466–22477. [\[CrossRef\]](#)
161. Wen, Y.; Kirchenbauer, J.; Geiping, J.; Goldstein, T. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv* **2023**, arXiv:2305.20030. [\[CrossRef\]](#)
162. Goyal, S.; Bunel, R.; Stimberg, F.; Stutz, D.; Ortiz-Jimenez, G.; Kouridi, C.; Vecerik, M.; Hayes, J.; Rebuffi, S.A.; Bernard, P.; et al. SynthID-Image: Image watermarking at internet scale. *arXiv* **2025**, arXiv:2510.09263. [\[CrossRef\]](#)

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.