

Article

Can World Foundation Models Generate Realistic Driving Videos? A Case Study on Pedestrian Crossing Scenarios

Cong Zhou ¹, Qian Lu ^{1,*}, Safraz Ahmed ¹, Olivier Haas ¹ and Vasile Palade ²

¹ Centre for Future Transport and Cities, Coventry University, Coventry CV1 5FB, UK; zhouc24@uni.coventry.ac.uk (C.Z.); ae5977@coventry.ac.uk (S.A.); o.haas@coventry.ac.uk (O.H.)

² Centre for Computational Science and Mathematical Modelling, Coventry University, Coventry CV1 5FB, UK; ab5839@coventry.ac.uk

* Correspondence: ad5271@coventry.ac.uk

Abstract

Autonomous vehicle (AV) technologies have advanced rapidly in recent years, driving an increasing demand for large-scale, high-quality annotated data. However, collecting and annotating real-world pedestrian video datasets is time-consuming, costly, and often insufficient to cover rare and safety-critical scenarios. Recent world foundation models have demonstrated impressive capabilities in generating realistic videos, yet their suitability for safety-critical autonomous driving applications remains largely unexplored. In this work, we investigate whether current world foundation models can generate driving scenarios that are sufficiently realistic and behaviourally consistent for autonomous driving research. We conduct a case study centred on pedestrian–vehicle interactions captured from ego-vehicle dashcam viewpoints, where subtle behavioural and geometric errors can have significant safety implications. To support this investigation, we develop SynPeDAS, an open research framework comprising a collection of synthetic pedestrian-interaction videos, a reusable generation pipeline for transforming real-world driving footage into synthetic scenarios, an automated evaluation suite, and downstream demonstration code. Through quantitative evaluation and structured human assessment, we identify several recurring failure modes, including dynamic misalignment, depth drift, and object persistence inconsistencies. More importantly, we find that commonly used evaluation metrics frequently exhibit ceiling effects and weak alignment with human judgement, limiting their ability to detect safety-critical behavioural errors. These findings indicate that, despite high perceptual realism at the frame level, current generative world models and existing evaluation methodologies remain insufficient for capturing physically grounded motion and task-critical semantics. Consequently, significant challenges remain before world model-generated videos can be considered reliable for safety-critical autonomous driving applications. SynPeDAS provides an open platform for systematically studying these challenges and developing improved generation and evaluation methods.



Academic Editors: Felipe Jiménez,
Jose Eugenio Naranjo and
Alfredo Valle

Received: 6 May 2026

Revised: 30 June 2026

Accepted: 7 July 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

Keywords: synthetic data; safe autonomous driving; world foundation model; pedestrian crossing; generative AI

1. Introduction

Autonomous vehicles (AVs) rely on robust perception systems to operate safely in complex urban environments. Among all road users, pedestrians present one of the most challenging perception and prediction problems due to their highly dynamic, context-dependent, and often unpredictable behaviours. Reliable pedestrian detection, tracking,

and intent classification are therefore critical components of high-level advanced driver assistance systems (ADAS).

Despite the availability of several public pedestrian datasets, they may be insufficient to train or evaluate ADAS across diverse pedestrian behaviours and environmental conditions. In addition, collecting and annotating real-world driving data remains costly, time-consuming, and logistically constrained. Rare and safety-critical scenarios are difficult to capture at a sufficient scale. As a result, perception models trained solely on real-world data may struggle to generalise to corner cases that are crucial for safety.

Recent advances in generative artificial intelligence, particularly world foundation models [1], have enabled the generation of realistic and temporally coherent videos from multimodal conditions. These developments have attracted increasing interest from the autonomous driving community [2], where synthetic data has the potential to complement traditional data collection pipelines and improve coverage of rare or safety-critical scenarios. A comparison between the traditional AV data collection process and a generative AI (GenAI)-based data collection process is shown in Figure 1. GenAI-based data collection enables rapid and on-demand generation of scenario-specific data, reducing the collection time required from weeks to hours while eliminating dependence on real-world conditions. However, despite rapid progress in visual realism, it remains unclear whether current world foundation models can faithfully reproduce the physical, behavioural, and semantic properties required by safety-critical autonomous driving applications.

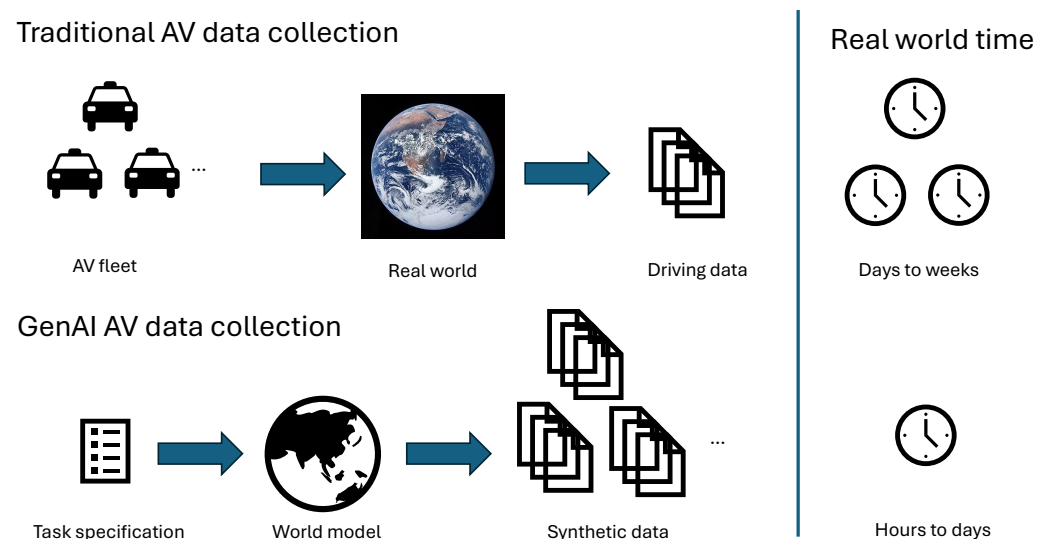


Figure 1. A comparison between the traditional AV data collection process and generative AI-based AV data collection. Traditional AV data collection relies on data recorded by vehicle fleets that are equipped with an AV sensor suite and data storage systems. It is also dependent on real-world conditions and the behaviours of other traffic participants. It might take days or even weeks to collect data that meets specific scenario requirements. In comparison, GenAI-based data generation is independent of real-world conditions. Plus, its data-gathering speed is significantly faster.

Existing research on synthetic data generation for autonomous driving has primarily investigated how to improve visual realism, scenario controllability, and dataset augmentation. Simulator-based approaches, such as CARLA, provide fine-grained control over scene composition and annotations, while learning-based generative methods aim to increase the photorealism and diversity of generated driving scenes. Recent driving-oriented world models further explore controllable video generation and scene synthesis as a means of expanding training and testing data for autonomous driving systems. Despite these advances, comparatively less attention has been paid to assessing the suitability of existing

generative-video evaluation metrics for autonomous driving. Many commonly adopted metrics were originally developed for general-purpose image or video generation and primarily assess appearance quality or coarse semantic alignment. Whether such metrics can capture safety-critical errors in generated driving scenarios remains an open question.

This motivates two fundamental research questions: First, to what extent can current world foundation models generate pedestrian-centric driving videos that are suitable for autonomous driving research? Second, do existing evaluation metrics adequately capture the safety-critical failures that emerge in generated driving videos?

To investigate these questions, we focus on pedestrian–vehicle interaction scenarios as a representative case study. Pedestrians are among the most vulnerable and behaviourally complex traffic participants in urban environments. Subtle errors in pedestrian motion, trajectory evolution, or interaction timing may have significant implications for downstream tasks, such as pedestrian detection, tracking, intention estimation, and risk assessment. Consequently, pedestrian-centric scenarios provide a particularly challenging benchmark for evaluating the applicability of world foundation models to autonomous driving.

In this paper, we introduce SynPeDAS, an open research framework consisting of a synthetic pedestrian video collection, a reproducible video generation pipeline, a quantitative evaluation suite, a human-based evaluation questionnaire, and downstream demonstration code. The framework is designed to be model-agnostic and can be integrated with different world foundation models that support multimodal conditioning. The rationale of this design choice is that the landscape of world foundation models is evolving rapidly, with new architectures, conditioning mechanisms, and generation capabilities emerging at a fast pace [3]. By combining controlled generation with systematic evaluation, SynPeDAS enables the study of generation quality, failure modes, and evaluation methodologies within a unified benchmark. To instantiate the framework and generate the synthetic videos used in our experiments, NVIDIA Cosmos Predict 2.5 [4] is used as the world foundation model, and real-world pedestrian scenes from the JAAD dataset [5] and the PIE dataset [6] are used as generation anchors.

To support this investigation, we use SynPeDAS to examine both the generation quality of world model-produced pedestrian videos and the suitability of existing evaluation protocols for this domain. The case study combines automatic video quality metrics with a structured human perceptual inspection, allowing us to assess not only visual realism and semantic alignment but also motion consistency, object persistence, and behaviour-level plausibility. By comparing metric-based results with human observations, this study further investigates whether commonly used evaluation methods are sensitive to failures that may be critical in autonomous driving scenarios.

The main contributions of this paper are as follows:

1. A systematic case study on the application of world foundation models to pedestrian-centric autonomous driving scenarios, using Cosmos Predict 2.5 as a representative driving-oriented model and providing empirical evidence of its current capabilities and limitations.
2. We present SynPeDAS, an open research framework comprising a synthetic pedestrian video collection, a reproducible generation pipeline, an evaluation suite, and downstream demonstration code.
3. Evaluation of the generated videos is provided, combining quantitative metrics and human perceptual assessment, which identify failure modes.
4. We demonstrate that existing evaluation metrics often fail to capture safety-critical behavioural errors and show weak alignment with human judgement, highlighting the need for domain-specific evaluation methodologies.

2. Related Work

2.1. Real-World Pedestrian Datasets for Autonomous Driving

Understanding pedestrian behaviour is a critical component of autonomous driving systems. Early datasets [7,8] primarily focused on pedestrian detection, providing bounding box annotations for object localisation. The JAAD [5] and the following PIE dataset [6] extend traditional detection benchmarks by incorporating behavioural and contextual annotations. They provide labels for pedestrian actions, environmental context, and road structure. This enriched annotation schema enables research on pedestrian intention prediction and interaction modelling.

Despite these advances, real-world pedestrian datasets have intrinsic limitations: Data collection and annotation are labour-intensive and costly. Safety-critical and rare events can be hard and not safe to collect. Moreover, naturalistic datasets are inherently biased toward frequently occurring scenarios, which results in limited coverage of edge cases that are essential for robust autonomous vehicle (AV) perception systems.

2.2. Synthetic Data for Autonomous Driving

Large-scale deployment of autonomous driving requires coverage of rare, safety-critical, and long-tail scenarios, which are difficult, expensive, or unsafe to capture in the real world. Therefore, a growing body of work explores different approaches to augment real-world driving data. Previous works can generally be categorised into simulator-based methods by using 3D engines and learning-based generative methods by using neural networks.

A notable simulator is CARLA [9], which is built on Unreal Engine 4. CARLA provides a flexible framework for evaluating autonomous driving systems and generating synthetic datasets through simulated sensors and dynamic environments. Due to its realism, extensibility, and open-source design, CARLA has become widely adopted in autonomous driving research. Despite these advances, datasets generated by CARLA or similar simulators often suffer from limitations in visual realism and scene diversity, as constructing complex scenes and detailed assets can be labour-intensive.

Simulator-based approaches have produced a range of synthetic driving datasets as concrete outputs of these generation methods. Early examples include the Fog Road Image Database (FRIDA), which contains synthetic urban road images under different fog conditions generated using the Sivic software [10]. Later datasets, such as SYNTHIA [11], Virtual KITTI [12], Synscapes [13], and Apollo Synthetic [14], substantially expanded the scale, visual diversity, annotation richness, and scenario coverage of synthetic autonomous-driving data. These datasets illustrate the practical value of simulation-based generation for producing controlled training and evaluation data while also highlighting the continuing trade-off between controllability, annotation availability, and visual-domain realism.

Another line of research relies on neural networks, particularly generative models. In 2017, ref. [15] proposed Pix2Pix, a conditional generative adversarial network (GAN) that can translate semantic labels into realistic images. This framework demonstrated the potential of learning-based approaches for large-scale semantic image generation. Numerous GAN-based methods have been proposed to synthesise realistic scenes from semantic inputs, such as [16–18].

Most recently, with the rise of large generative models, a few experiments [19] have been conducted to explore the possibility of building a video generator tailored to driving contexts. A comprehensive survey on synthetic data generation methods can be found in [20]. It highlights the trade-offs between controllability and realism: physics-based simulators offer precise annotations and scenario control but often exhibit domain gaps, whereas data-driven generative methods achieve higher photorealism but provide weaker guarantees of physical consistency and semantic correctness.

Our work builds on this line of research but shifts the focus from general synthetic driving data generation to a targeted case study of world model-generated pedestrian-interaction videos. We examine whether current generative world models can produce pedestrian-focused driving videos that preserve scene semantics, temporal coherence, and behaviour-level plausibility. This focus complements existing synthetic data studies by emphasising the evaluation of generated video quality and failure modes in a safety-relevant pedestrian scenario.

2.3. World Foundation Models for Video Generation

Modelling the physical world itself is not a new idea. In traditional automation, control, and robotics, mathematical and data-based models have long been used for planning, prediction, and control. These models may be derived from first-principles laws, such as kinematic or dynamic equations, or parameterised using system identification from observed input–output data. Depending on the availability of physical knowledge and data, such approaches can range from white-box models with physically interpretable parameters, to grey-box models that combine mechanistic structure with data-driven estimation, and black-box models whose parameters may not have direct physical meaning [21,22]. However, these conventional approaches are usually designed for specific systems and operate in relatively low-dimensional state spaces. As a result, they often struggle to generalise across different systems or environments, limiting their reuse for new tasks.

The concept of world models was first introduced in [23], which proposed a model-based reinforcement learning method that learns an internal representation of the environment using neural networks. The world model and policy network are trained using interaction records between an agent and the environment. The world model, which uses a recurrent neural network as its backbone, can predict future states based on current states and inputs. A sufficiently accurate representation of the environment allows systems to anticipate future outcomes and select actions that have better expected rewards. Follow-up research, such as the Dreamer series [24–26] and Td-mpc2 [27], explored the modelling of more complex environments.

More recent research treats world models as generative models operating in the visual observation space. These approaches commonly use conditional video generation models, such as action-to-video or text-to-video approaches. Models may follow an autoregressive formulation [28,29] or adopt diffusion-based techniques [30]. Another emerging direction is generative simulation, which integrates generative AI models with physical simulation engines to better capture real-world dynamics [31,32]. More recently, V-JEPA 2-AC [33] demonstrates that learned world representations through self-supervised video pretraining can be leveraged for zero-shot robotic planning.

Among these developments, NVIDIA’s Cosmos [4] stands out as the only world foundation model that has been explicitly post-trained on large-scale real-world driving data for autonomous driving simulation. It also provides a dedicated post-training pipeline fine-tuned on driving-specific data, enabling physics-aware video generation and photorealistic scene transfer for autonomous driving scenarios. This domain-specific focus on driving makes Cosmos the most suitable choice as the video generator for SynPeDAS.

2.4. Evaluation of Generative World Models

Alongside the development of generative world models, evaluation has become an important research direction. The evaluation methods discussed in the literature can be categorised into three high-level types: traditional quantitative metrics, model-based evaluation frameworks leveraging Large Language Models (LLMs) and Visual Language Models (VLMs), and specialised expert models.

Quantitative metrics use established measures, including the Inception Score (IS), Fréchet Inception Distance (FID), and Fréchet Video Distance (FVD) to calculate general video quality [19]. Deep feature embeddings derived from pre-trained models, such as CLIP [34], are utilised to measure video–text alignment. In addition, ref. [35] adopts measures from classical computer vision, including photometric consistency calculated using the Average End-Point Error (AEPE) from optical flow estimation and assessing 3D consistency by estimating dense pixel-wise depth. However, a significant drawback is that the evaluation results frequently exhibit substantial misalignment with human preferences [36]. These metrics typically offer limited insight into complex capabilities, e.g., alignment with physical rules.

Model-based approaches leverage the reasoning capabilities of Multimodal Large Language Models (MLLMs) as an alternative to hand-crafted metrics and costly human evaluation. Generalist VLMs and LLMs are used to perform structured assessment over multimodal inputs (e.g., video frames and text), as demonstrated in [36,37]. These approaches offer strong scalability and flexibility, enabling the evaluation of complex attributes, such as temporal consistency and semantic alignment, without requiring extensive human annotation.

Nevertheless, their reliability as proxies for human judgement is limited. LLM-based evaluators exhibit systematic misalignment with human preferences, particularly when producing scalar scores [38]. This misalignment stems from differences in implicit evaluation criteria and sensitivity to prompting, leading to inconsistent or biased judgements. Consequently, while model-based evaluators reduce cost and improve coverage, they should be treated as approximate and imperfect substitutes for human evaluation, often requiring calibration or pairwise comparison schemes to improve alignment.

Specialised Expert Detectors are models, fine-tuned using human-annotated data, to assess evaluation dimensions where general LLMs/VLMs are known to struggle. Ref. [39] collected 67,000 human labels and fine-tuned a lightweight Visual Language Model with 2 billion parameters to predict human preference with a 9.9% lower error rate than GPT-4o. Ref. [37] trained a vision transformer-based specialist using human-annotated generated samples to detect structural anomalies in human figures. Nevertheless, developing these specialists requires extensive manual annotation and data curation.

3. SynPeDAS Framework

3.1. Framework Overview

SynPeDAS is designed as an integrated framework for investigating the capabilities and limitations of world foundation models in pedestrian–vehicle interaction scenarios. We focus on pedestrian crossing events because these scenarios involve complex interactions between pedestrians, vehicles, road geometry, and the environment. Successful generation of pedestrian crossing scenes requires not only a realistic appearance but also plausible motion evolution, stable geometric relationships, and behaviour that remains consistent with both scene context and textual conditioning. As a result, pedestrian crossings provide a natural benchmark for assessing the strengths and limitations of current world foundation models.

The framework consists of three core components: real-world scenario sourcing, a video generation pipeline, and an evaluation suite. In addition, SynPeDAS includes a pedestrian detection demonstration as downstream demonstration code, which is presented in Section 6 after the evaluation sections. Real-world scenario sourcing identifies and prepares safety-relevant pedestrian clips from existing driving datasets. The video generation pipeline transforms these real clips into synthetic driving videos by combining source observations with structured textual descriptions. The evaluation suite assesses generated

videos using both automatic metrics and human perceptual judgement. Finally, the detection demonstration illustrates how generated data can be incorporated into a downstream pedestrian perception workflow. The whole SynPeDAS framework (version 1.0), as well as the video collection we generated, can be accessed through our GitHub repository (<https://github.com/levilovearch/SynPeDAS> (accessed on 6 July 2026)).

3.2. Real-World Scenario Sourcing

We use real pedestrian–vehicle interaction scenes from the JAAD dataset and the PIE dataset as generation anchors. Both datasets provide urban driving videos with annotated pedestrian behaviours, making them suitable for selecting safety-relevant crossing events. To ensure that the generated data focuses on safety-critical moments, we extract temporally localised clips around semantically meaningful events, such as pedestrians approaching a curb, pedestrians initiating crossing, and vehicle yielding dynamics.

These keyframes define temporal anchors from which short video segments are extracted. Each segment is trimmed to remove irrelevant pre- and post-event frames, ensuring that the clip concentrates on the target pedestrian interaction. This temporal focusing limits semantically irrelevant context, promoting downstream captioning to remain grounded in the pedestrian crossing event, and therefore improves generation controllability. Following the filtering conditions detailed in Appendix A, we identified 330 pedestrian crossing scenarios. These filtering conditions are intentionally designed to be strict, to prioritise scenario relevance and behavioural clarity. Users can adapt them according to their own scenario definitions, trading off the number of matched scenarios against the degree of conformity to the target scenario.

3.3. Video Generation Pipeline

The pipeline transforms sourced real-world pedestrian scenes into synthetic driving videos while preserving scenario semantics and improving controllability. Our approach leverages multimodal-conditioned generation, using both initial real-video observations and structured textual descriptions as generation conditions. An overview of the pipeline is illustrated in Figure 2. While NVIDIA Cosmos Predict 2.5 is used as the world foundation model in our experiments, the pipeline is designed to be model-agnostic and can be adapted to other video generative models that support multimodal conditioning.

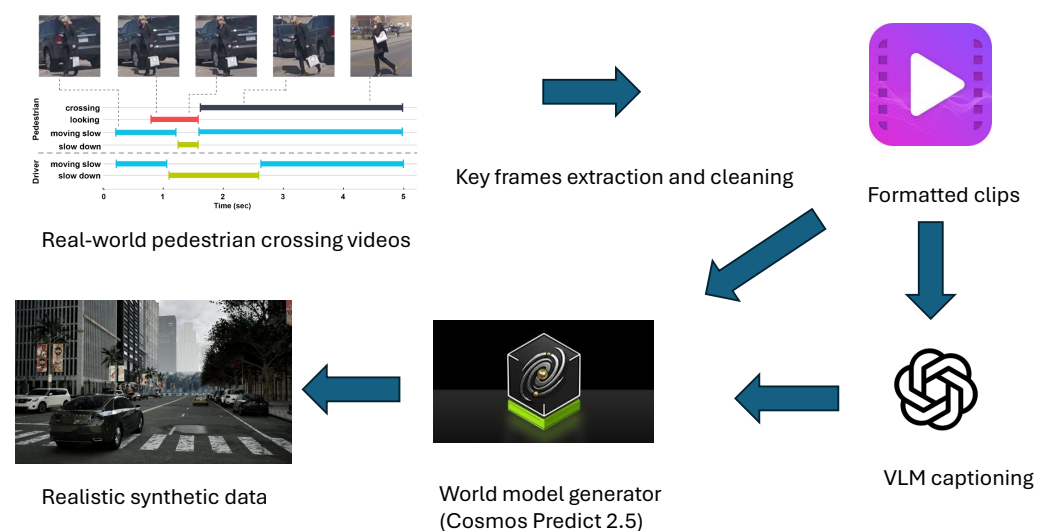


Figure 2. Layout of the scenario sourcing and data generation pipeline used in this work. Arrows indicate the flow of information between processing stages. It consists of three main stages: (1) keyframe extraction and clip selection, (2) generating structured scene descriptions via a vision language model, and (3) synthetic video generation using Cosmos Predict.

3.3.1. Structured Scene Description with VLM

Each selected clip is analysed using a Vision Language Model (VLM) to produce a detailed scene description. Considering captioning quality and cost efficiency, we use Google Gemini 2.5 Flash [40] to generate all video captions. Rather than relying on the VLM to produce free-form descriptions without guidance, we carefully designed the prompt to ensure comprehensive coverage of scene elements that are critical to autonomous driving scene definition. To this end, we adopt the 6-Layer Model [41] as a conceptual framework, directing the VLM to address road infrastructure, roadside structures, dynamic temporal modifications, scene objects, daytime and weather conditions, and communication elements, such as traffic signals. In addition, the prompt incorporates cinematographic language conventions, instructing the VLM to describe camera composition, lighting setups, and material-level sensory details, such as textures, reflections, and shadows. An example of the generated caption is shown in Figure 3.

A wide-angle perspective from a vehicle's dashboard captures a bustling commercial thoroughfare under an overcast sky, casting a soft, diffused natural light across the scene. The paved asphalt road stretches forward, flanked on the right by a long, multi-story building featuring warm brickwork and modern green paneling, with large display windows reflecting the muted daylight. Several pedestrians traverse the sidewalk adjacent to the building, while others begin to step across the access road, one figure walking away from the camera's view. To the left, a busy parking lot is filled with numerous parked cars of various models, bathed in the same even light, beneath a tall streetlamp with brightly glowing fixtures. In the distance, a distinctive, tall cylindrical high-rise building stands prominently against the cloudy horizon, anchoring the urban backdrop. The steady shot captures the continuous, natural movement of people within this everyday suburban retail environment.

● Scene Content (6LM: Road, Structures, Objects, Infrastructure) ● Environmental Conditions (6LM: Weather, Lighting, Temporal)

● Cinematographic Language ● Materials & Sensory Details

Figure 3. An example of a VLM-generated scene description for a pedestrian video clip. The colour-coded annotations illustrate the coverage of our prompt design. Highlighted phrases indicate four categories of scene information: Scene Content—elements aligned with the 6-Layer Model (6LM) framework, including road infrastructure, roadside structures, pedestrians, vehicles, and traffic fixtures; environmental conditions—weather, lighting, and temporal dynamics; cinematographic language—camera perspective, composition, and shot attributes; and Materials and Sensory Details—surface textures, reflections, and lighting properties.

A notable empirical finding is that structured formats, such as bullet points or listings in captions, severely degrade the quality of generated videos; the prompt therefore explicitly prohibits the use of lists and enforces continuous descriptive prose of fewer than 250 words. The full prompt can be found in Appendix B. By guiding the VLM through domain-specific heuristics, we ensure that the resulting descriptions are both physically grounded and semantically rich, improving alignment with the conditioning requirements of world foundation models.

3.3.2. Synthetic Video Generation with Cosmos Predict

NVIDIA Cosmos Predict 2.5 is a diffusion transformer-based video generator capable of synthesising future frames conditioned on multimodal inputs, including textual prompts, depth maps, segmentation masks, and initial observation frames. These conditioning modalities can be freely combined within a single generation pass. In our setting, the first five frames of each real video clip serve as the conditioning observation, while the textual prompt is derived from the procedure described in Section 3.3.1. Generation is performed fully automatically without any manual post-editing. Each synthesised clip is rendered at 720p resolution with a duration of 5 s. This duration was selected to balance visual quality with the preservation of meaningful pedestrian behaviour semantics. All gen-

erator hyperparameters were kept at their default values as recommended by NVIDIA [42]. The generation process was executed on a cloud-hosted single H200 GPU instance. Synthesising the entire dataset required approximately 25 h under this hardware configuration, and the instance cost is around 300 USD.

Figure 4 presents a compelling example of a synthesised video, showing visually natural environments and plausible pedestrian motion. A notable difference from the real clip lies in the pedestrian distribution: the real sequence captures a single group passing through, whereas the generated version depicts two distinct groups. Additionally, although the textual prompt (see Appendix C.1) specifies that the ego vehicle should turn into the alley on the right, the generated vehicle remains stationary, yielding to the denser pedestrian crowd. This deviation from the prompt is not a failure of the model. It reflects a physically plausible traffic interaction where the vehicle prioritises collision avoidance over route adherence. This suggests that Cosmos Predict 2.5 learns to reconcile textual conditioning with scene-level physical consistency, favouring coherent behaviour when conflicting cues are present.

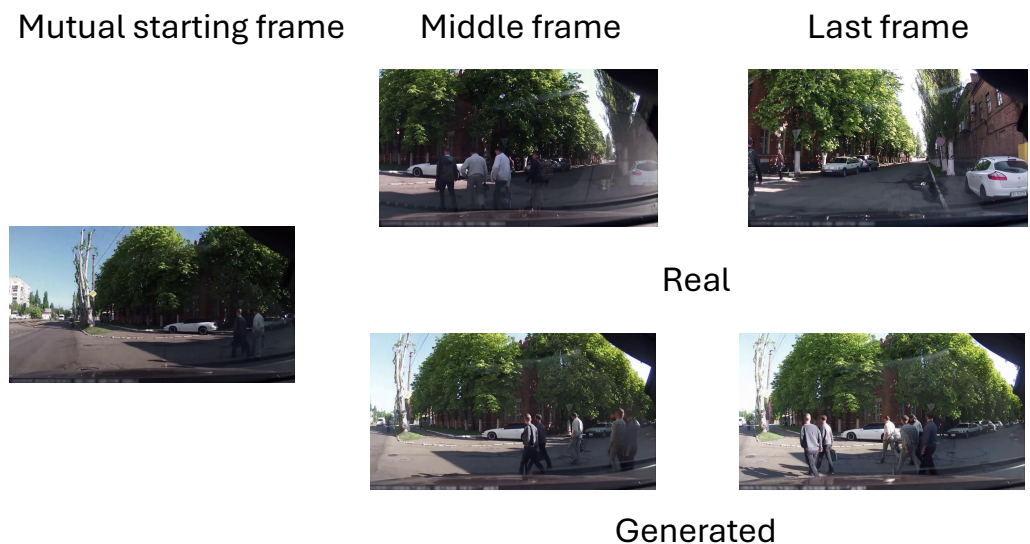


Figure 4. Visual comparison of one representative clip. Top row: real video frames at $t = 0, 3, 5$ s. Bottom row: corresponding generated frames conditioned on the first 5 real frames and the textual prompt. The prompt we used is in Appendix C.1.

3.4. Video Collection Format

Using this pipeline, we generate 330 synthetic pedestrian driving videos with a total duration of 1650 s at 720p resolution. These synthetic clips preserve the semantic structure of real scenarios from the JAAD and PIE dataset while introducing controlled variations. Our dataset can also be accessed through our GitHub repository.

The dataset is organised into three primary subfolders: (1) generated videos, (2) corresponding original video clips, and (3) generation prompts used for video synthesis. The generated videos folder contains all synthetic clips produced by the world model-based generator. The original video clips folder stores the real-world source videos from which scenarios were selected and grounded. The generation prompts folder includes the structured textual descriptions used to condition the video generation process.

In addition, the evaluation results for each generated video are provided in a structured CSV file. This file contains the computed scores for all selected evaluation metrics listed in Section 4, enabling researchers to filter and select videos based on specific quality thresholds or metric combinations. By providing both the generated content and its associated evaluation scores, SynPeDAS facilitates reproducible research and allows for

systematic exploration of the relationship between generation quality and downstream task performance.

3.5. Evaluation Suite

The SynPeDAS evaluation suite serves as a video quality assessment component for the framework. Its metric set covers the most relevant and currently valuable dimensions available for synthetic driving-video assessment, including semantic alignment, geometric and photometric consistency, perceptual quality, pedestrian anatomy, temporal smoothness, and distributional similarity. Together, these metrics support a broad assessment of whether generated videos are visually plausible, temporally stable, structurally coherent, and semantically aligned with their conditioning inputs. The detailed metric definitions and quantitative results are provided in Section 4; the human perceptual validation is discussed in Section 5.

4. Quantitative Evaluation

To assess the realism, consistency, and task relevance of SynPeDAS, we conduct both a quantitative evaluation using established video quality metrics and a structured human perceptual study. Our evaluation focuses on three aspects critical to ADAS perception: visual quality, spatial-temporal consistency, and behavioural plausibility.

It should be noted that the quantitative evaluation in this work is intended as an intra-framework diagnostic analysis rather than a cross-method benchmark against other synthetic data generation paradigms. Simulator-based approaches, GAN-based generative methods, and world foundation model-based video generation differ substantially in their conditioning inputs, controllability, annotation availability, output modality, and target use cases. Therefore, direct metric comparison across these paradigms may be misleading without a carefully normalised benchmark protocol. Our evaluation focuses on characterising the quality, failure modes, and metric limitations of the generated videos produced by the proposed SynPeDAS pipeline. A systematic benchmark comparing world foundation models with simulator-based and earlier generative approaches is left for future work.

4.1. Selection of Metrics

Existing metrics for synthetic videos capture both low-level visual fidelity and high-level semantic coherence. In this work, we select seven metrics, described in the following paragraphs, which can be categorised into four key dimensions: content alignment, geometric and photometric consistency, human anatomy and motion, and distributional distance.

Content alignment measures the semantic consistency between the generated video and its conditioning input (e.g., text prompt or start frame). The FLEUR metric [43] was originally designed for reference-free image captioning. FLEUR leverages a Large Multi-modal Model (LMM), specifically LLaVA [44], to evaluate the alignment between visual content and text without relying on human-written reference captions.

We adapted this image-level metric for video by uniformly sampling frames from the generated clip, following [35]. For each frame, we generate a caption using the LMM and compute its alignment score with the conditioning text. The final content alignment score for the video is the mean of these frame-level scores. This aggregation provides a holistic measure of how well the generated video maintains semantic fidelity to the input condition throughout its duration.

Geometric consistency evaluates whether the 3D structure and the geometry of the objects remain coherent throughout the video. Following [35], we adopted DROID-SLAM [45], a SLAM-based approach that jointly estimates pixel-wise depth maps and relative camera poses for the sequence. This allows us to quantify the geometric stability

via the reprojection error, measuring the discrepancy between observed pixel movements and those implied by the estimated 3D structure.

Photometric consistency measures the stability of pixel-level brightness and colour across consecutive frames. Inconsistent appearance may indicate quality issues, such as texture flickering. Following [35], we evaluate this by warping the previous frame to the current frame using the estimated optical flow and calculating the mean-squared error between the warped frame and the actual current frame. Lower errors indicate higher consistency.

Subjective quality aims to capture the human-perceived realism and technical fidelity of the generated video. Following the methodology of [35], who validated a combination of aesthetic and quality metrics against human judgements, we utilise CLIP-IQA+ as our primary indicator of perceptual quality.

CLIP-IQA+ operates by leveraging a pre-trained CLIP model [46] with two learnable continuous text prompts: a positive prompt T_1 representing “high-quality” and a negative prompt T_0 representing “low-quality.” For a given video frame I , we extract its image embedding using the frozen CLIP image encoder. The metric computes the cosine similarity between the image embedding and both text prompt embeddings. To obtain the video-level subjective quality score, we compute the mean Frame-IQA score across T uniformly sampled frames.

Human anatomy evaluates the structural integrity and physiological plausibility of human figures within the generated video. We adopt the anomaly detection framework from [37], utilising only the full-body anomaly detection module to align with the pedestrian-focused nature of our dataset.

Motion smoothness assesses the temporal continuity and naturalness of motion within the generated video. Following [37], we employ the AMT model [47] for video frame interpolation to estimate intermediate frames. Motion smoothness is then quantified by measuring the Structural Similarity Index (SSIM) between the original frames and their interpolated counterparts. Higher SSIM scores indicate greater structural consistency, reflecting smoother motion with reduced flickering or artefacts. The final metric is computed as the mean SSIM across all interpolated frames in the sequence.

CMMD evaluates the distributional similarity between generated and real video frames, serving as a robust proxy for perceptual quality. Addressing the limitations of the Fréchet Inception Distance (FID) [48], such as bias and reliance on normality assumptions, CMMD computes the unbiased Maximum Mean Discrepancy (MMD) in the CLIP embedding space [49]. Using the semantic richness of CLIP and avoiding parametric assumptions, CMMD can provide a more reliable and sample-efficient assessment of generation quality that aligns closely with human judgement. We report the CMMD score between our generated frames and the ground-truth test set to quantify overall fidelity.

4.2. Evaluation Results and Discussion

As shown in Figure 5 and Table 1, the results indicate strong performance in prompt adherence and perceptual realism. The dataset achieves a mean content alignment score of 0.98, demonstrating that the generated videos closely follow the intended textual descriptions. The subjective quality score reaches 0.86, suggesting high overall visual fidelity and minimal perceptual artefacts. Human anatomy consistency is particularly strong (mean = 0.99), indicating that the pedestrian body structure is well preserved without noticeable distortions, which is an essential property for downstream tasks, such as pedestrian detection and tracking.

Several metrics exhibit ceiling effects: their scores cluster near the upper bound across all generated samples, offering limited discriminative power. This aligns with the previous literature’s claim [36,39] on the limitations of quantitative metrics. Content

alignment, human anatomy, and motion smoothness are tightly concentrated near the upper bound, indicating consistently high performance across samples. While this reflects strong generation quality, it also indicates the limited discriminative power of these metrics for fine-grained ranking or filtering. In contrast, “subjective quality” shows moderate variability, suggesting that it remains sensitive to perceptual differences across clips.

Table 1. Evaluation distribution status of all seven metrics we adopted to evaluate generated video quality.

Metric	Mean	Median	Std
Photometric Consistency	0.45	0.53	0.33
3D Consistency	0.49	0.51	0.25
Subjective Quality	0.86	0.85	0.09
Content Alignment	0.98	0.98	0.02
Human Anatomy	0.98	1.00	0.05
Motion Smoothness	0.99	0.99	0.01
Cmmd	0.41	0.45	0.19

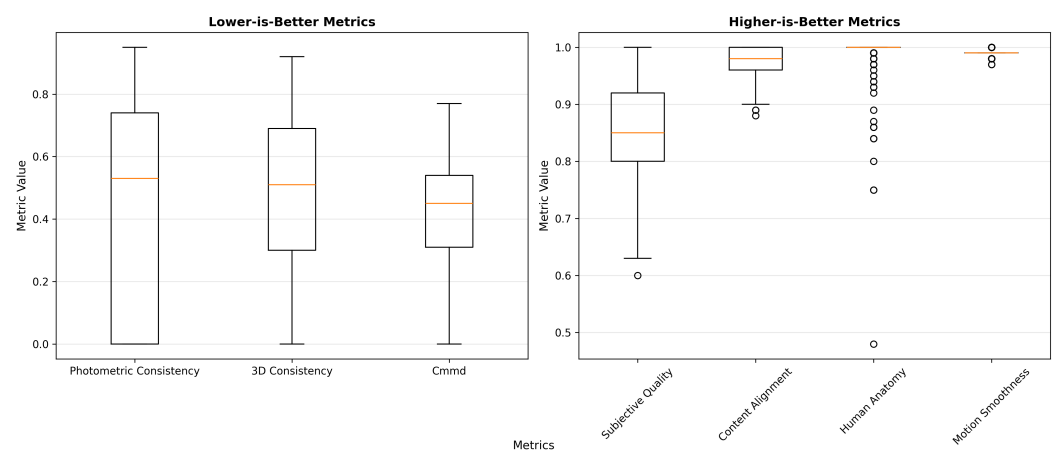


Figure 5. Boxplot of evaluation metrics across all generated videos. The central line indicates the median, while the box represents the interquartile range (IQR) between the 25th and 75th percentiles. Whiskers denote the range of non-outlier values. Metrics such as content alignment, human anatomy, and motion smoothness exhibit strong ceiling effects with low variance, whereas photometric consistency and 3D consistency show substantially higher variability, indicating their greater sensitivity in capturing quality differences across generated videos.

Greater variability is observed in consistency-related metrics. The 3D consistency score (mean = 0.49, SD = 0.25) spans a wide range, indicating that geometric coherence is not uniformly preserved across all samples. Similarly, CMMD (mean = 0.41, SD = 0.19) suggests variability in modelling larger-scale, depth-aware motion. These results imply that while individual frames are often visually convincing, structural consistency over time, particularly under motion, remains a key challenge.

Photometric consistency (mean = 0.45, SD = 0.33) exhibits a large dispersion; however, this metric is of limited suitability for autonomous driving scenarios. Photometric consistency metrics are typically designed under the assumption of relatively stable illumination conditions, as commonly found in indoor or controlled environments. In contrast, driving scenes inherently involve rapid and frequent lighting changes, such as transitions through tunnels, occlusions from surrounding buildings, or disturbances caused by rain and windshield wipers. As a result, low photometric consistency scores in this setting primarily reflect natural environmental variability rather than deficiencies in generation quality, and should therefore be interpreted with caution.

Temporal coherence remains highly stable, with motion smoothness achieving a mean score of 0.990. This indicates minimal frame-to-frame jitter or flickering and suggests that the underlying generative model, NVIDIA Cosmos Predict 2.5, provides strong short-term temporal consistency.

Overall, the results suggest that SynPeDAS videos achieve high perceptual realism and strong semantic alignment, while exhibiting moderate limitations in geometric consistency under motion. Consistency-related factors, particularly geometric coherence, remain the primary source of variability in the dataset. These findings also highlight the inadequacy of traditional evaluation metrics for synthetic driving data. While conventional metrics (e.g., FID) and embedding-based scores (e.g., CLIP) effectively capture low-level visual quality and coarse semantic alignment, they often fail to assess physical realism and task-critical fidelity. A generated driving video clip may achieve high CLIP scores while violating fundamental traffic constraints or producing subtle but safety-critical errors (e.g., incorrect turning direction) that are diluted in global similarity measures. Furthermore, standard metrics such as photometric consistency may not be suitable for dynamic driving environments, where changes in natural illumination are often misinterpreted as generation artefacts. Consequently, there is a pressing need for domain-specific metrics that prioritise physical plausibility and fine-grained semantic correctness over generic visual fidelity.

5. Human Perceptual Inspection for Failure Pattern Identification

To complement the automatic metric-based evaluation, we conducted a small-scale human perceptual inspection of the generated videos. The purpose of this inspection was not to provide a statistically representative assessment of the full SynPeDAS video collection. Instead, it was designed as a structured diagnostic procedure to identify recurring perceptual and behavioural failure patterns that may not be sufficiently captured by automatic metrics. The resulting descriptive human inspection scores are summarised in Table 2.

Table 2. Descriptive human inspection results on 50 randomly sampled video clips using a 1–4 Likert scale across scenario fidelity, environmental correctness, and motion plausibility.

Category	Criterion	Mean Score (1–4)
Environmental and Scene Fidelity	Lighting accuracy	3.57
	Weather correctness	3.47
	Intensity realism	3.40
	Road category clarity	3.37
	Structural consistency (bridges, sidewalks)	3.37
Scene Composition and Object Presence	Lane markings consistency	3.13
	Traffic sign correctness	2.77
	Key object presence	2.63
Behavioural Plausibility	Specific maneuver correctness	2.79
	Physical plausibility	2.90
Caption adherence	Object movement alignment with caption	2.03
	Overall caption correspondence	2.57

The inspection involved nine volunteers aged between 25 and 40, recruited from Asian and European backgrounds. All participants reported normal or corrected-to-normal vision and possessed general familiarity with urban driving scenarios. Each participant evaluated approximately 10 randomly sampled generated videos. Across all responses, the inspected subset covered 50 unique video clips, with overlapping ratings for a subset of clips. Given the limited number of participants and clips, the results reported in this section should be

interpreted as descriptive evidence for failure pattern analysis rather than as generalisable human preference statistics.

Participants were instructed to watch the AI-generated driving videos alongside their corresponding text prompts. Real-world driving footage was used as a contextual reference to help participants judge whether the generated scenarios remained plausible with respect to the original driving context. Participants rated each video on a 1–4 Likert scale across four dimensions: environmental fidelity, object consistency, behavioural plausibility, and caption adherence. The primary objectives were to identify perceptual artefacts, object-level inconsistencies, and motion-related failures, and to assess whether the generated visual content remained aligned with the textual prompt. The detailed questionnaire and rating criteria are provided in our GitHub repository.

The descriptive ratings suggest that environmental and scene-level attributes were less problematic than motion- and behaviour-related attributes in the inspected subset. Lighting accuracy receives the highest mean score of 3.57, followed by weather correctness at 3.47 and intensity realism at 3.40. Road category clarity and structural consistency, including bridges and sidewalks, both reach 3.37. These observations suggest that the generated clips often preserve plausible environmental appearance and background structure. This is consistent with the design of our generation pipeline, where real-world frames are used as visual anchors for generation. However, these ratings should not be interpreted as evidence that the entire video collection is uniformly realistic, but rather as an indication that environmental fidelity was not the dominant failure source in the inspected samples.

Object-level consistency shows moderately strong performance. Lane marking consistency achieves a mean score of 3.13, while traffic sign correctness scores 2.77. The presence and consistency of key scene objects receive a mean rating of 2.63. Although these scores indicate generally acceptable realism, they reveal occasional omissions or inconsistencies in secondary objects. Figure 6 captures one such instance: between consecutive frames at $t = 3$ and $t = 4$, stop signs at two marked locations abruptly transform into traffic lights—a sudden swap in object identity that violates temporal coherence. Such frame-to-frame identity flips are particularly detrimental for downstream tasks that rely on consistent object semantics, such as traffic sign recognition or scene understanding benchmarks.



Figure 6. Generated frames at $t = 3$ (a) and $t = 4$ (b). Yellow arrows indicate the locations where stop signs in $t = 3$ are replaced by traffic lights in $t = 4$, demonstrating sudden object identity inconsistency across consecutive frames.

Motion and behavioural plausibility present greater challenges. The lowest average rating is observed for object movement alignment with the caption, which receives a mean score of 2.03. Specific maneuverer correctness scores 2.79, overall caption correspondence 2.57, and physical plausibility 2.90. These results suggest that while motion appears visually smooth, it is not always semantically aligned with the intended scenario.

Figure 7 provides a concrete example: in the real clip, a pedestrian crosses the street briskly without interruption, whereas in the generated counterpart, the same pedestrian pauses directly in front of the oncoming vehicle and turns to look back toward the direction of origin. Although the generated motion, which is a person stopping and looking back, is

physically natural, it diverges from the prompts, which state “A pedestrian, dressed in dark attire, begins to cross the street from the left sidewalk, moving purposefully through the stopped traffic lanes toward the right side of the frame”. Such discrepancies in trajectory, stopping behaviour, and interaction timing are perceptually noticeable and contribute to the relatively low scores in motion alignment.

In summary, the human perceptual inspection provides diagnostic evidence that the most salient failures are not primarily related to environmental appearance but to object persistence, behavioural alignment, and prompt-conditioned motion grounding. These observations motivate the failure mode analysis in Section 7, where we jointly examine human observations and automatic metric results to characterise recurring limitations of generated pedestrian driving videos.

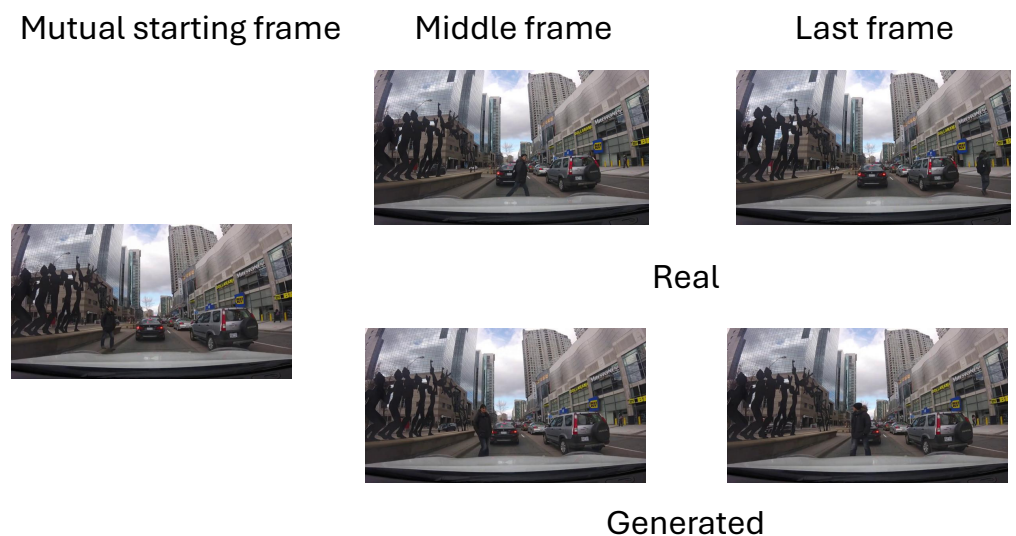


Figure 7. Real (top) and generated (bottom) frames at $t = 1, 3, 5$ s, illustrating a pedestrian crossing scenario. The prompts used for generation are in Appendix C.2. In the real clip, the pedestrian crosses briskly without stopping. In the generated clip, the pedestrian pauses directly in front of the vehicle and looks back toward the direction of origin.

6. Detection Demo

To provide an initial downstream assessment of SynPeDAS, we conduct a pedestrian detection demo using synthetic data generated by our pipeline. Pedestrian detection is selected because it is a fundamental perception task in autonomous driving and is primarily driven by frame-level appearance cues. It therefore provides a suitable test bed for examining whether synthetic pedestrian videos can support downstream model training. The goal of this demo is not to claim that SynPeDAS is immediately applicable to all safety-critical ADAS tasks. Instead, we use detection as a controlled downstream probe to investigate whether generated data becomes more useful when it is curated by the proposed evaluation suite. Other downstream applications, such as tracking, segmentation, trajectory prediction, behaviour understanding, and scenario-based safety analysis, may require different annotation schemes and data-selection strategies, and are left for future work.

We use the JAAD dataset as both the real-data source and the evaluation benchmark, following the split defined in [5]. Synthetic clips are used only as training augmentation, while the evaluation set is kept fixed across all experimental configurations. The target detector is a faster R-CNN model implemented in Detectron2 [50]. We fine-tune the model for 300 iterations with a batch size of 32 and a learning rate of 0.00025.

Since the generated videos do not contain manually annotated frame-level bounding boxes, we adopt a coarse-grained automatic pseudo-labelling strategy. Specifically, we use

YOLOv11L [51] as an off-the-shelf pedestrian detector to generate bounding box annotations for pedestrians in the synthetic videos. This design avoids the high cost of manual frame-level annotation and provides a practical way to use generated videos for detector fine-tuning. However, we note that pseudo labels may introduce additional annotation noise, and this demo should therefore be interpreted as a practical downstream experiment rather than a fully controlled annotation-quality study.

We compare four experimental settings. The first setting is the pre-trained baseline without fine-tuning. The second setting fine-tunes the baseline model using real-world JAAD images only. The third setting fine-tunes the model using both real-world data and all generated data, representing a naive synthetic augmentation strategy. The fourth setting fine-tunes the model using real-world data and quality-filtered generated data. In this setting, generated clips are selected according to the quality score computed by the evaluation suite. Each metric is first normalised to the range $[0, 1]$, with directionality correction applied so that higher values consistently indicate better quality. The per-video quality score is then computed as the unweighted arithmetic mean of all normalised metric scores. We retain only generated clips with a quality score above 0.80. After frame extraction, this results in approximately 5000 synthetic training images.

To make the comparison between real-only training and quality-filtered synthetic augmentation more balanced, we use approximately 5000 real images in both the real-only setting and the quality-filtered synthetic setting. This produces a roughly equal number of real and selected synthetic images in the final quality-filtered training set. These settings allow us to compare three cases: fine-tuning with real data only, naively adding all generated samples, and adding a smaller but quality-filtered subset of generated samples.

The results are shown in Table 3. Fine-tuning with real-world JAAD data substantially improves AP from 49.72% to 75.55%, confirming that the detector benefits from task-specific fine-tuning on the target data distribution. However, naively adding all generated samples reduces AP to 70.86%, which is lower than the real-only fine-tuning result. This indicates that simply increasing the number of training samples with unfiltered synthetic data does not necessarily improve downstream performance. Low-quality generated samples, pseudo-label noise, or domain mismatches may introduce harmful supervision signals during fine-tuning.

Table 3. Pedestrian detection results across real-only fine-tuning, unfiltered synthetic augmentation, and quality-filtered synthetic augmentation settings.

Setting	No. Samples	AP
Baseline	-	49.72%
Fine-tuned (Ft.) by real-world data	5114	75.55%
Ft. by real-world data + generated data	22,689	70.86%
Ft. by real-world data + quality-filtered generated data	10,228	77.91%

By contrast, using only quality-filtered generated samples achieves the best AP of 77.91%, improving over the real-only fine-tuning setting by 2.36 percentage points. Although this result should be interpreted as an initial demonstration rather than a comprehensive downstream benchmark, it suggests that evaluation-guided data curation can improve the practical utility of synthetic data. In particular, the comparison between using all generated samples and using a smaller quality-filtered subset shows that synthetic data quality can be more important than synthetic data quantity alone.

This demo provides preliminary evidence that SynPeDAS can support perception tasks, such as pedestrian detection, when generated samples are selected through a quality assessment mechanism. At the same time, the result also highlights the limitations of naive

synthetic augmentation. The quality score used in this experiment is based on a simple unweighted average of normalised metrics. Future work could investigate task-specific weighting strategies, additional safety-oriented metrics, and more accurate annotation pipelines to further improve the effectiveness of synthetic data for downstream autonomous driving tasks.

7. Failure Mode Analysis

Joint analysis of automatic metrics and the diagnostic human inspection reveals three dominant failure modes: The most prevalent issue is dynamic misalignment. In these cases, pedestrian trajectories subtly deviate from the scene geometry or the real-world walking pace. For example, as shown in Figure 8, a woman in yellow clothing in the real-world video nearly completes her crosswalk traversal within approximately two seconds ($t = 3$ s to $t = 5$ s), whereas in the generated video, the same pedestrian moves noticeably slower and remains near the right side of the road even by the end of the clip. Although the motion smoothness score remains high in such cases, the semantic consistency between motion and scene context is occasionally compromised. This further highlights the limited discriminative power of traditional motion smoothness metrics for fine-grained ranking.

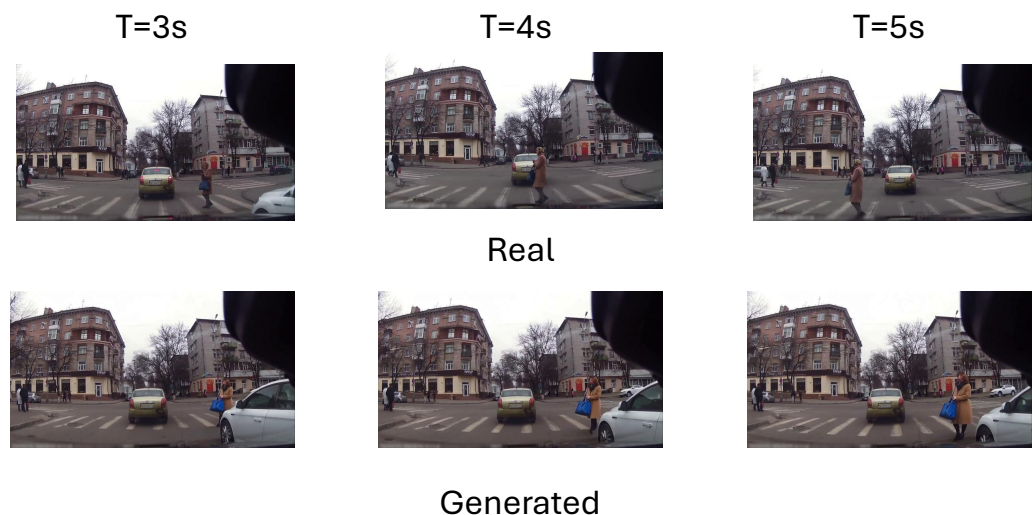


Figure 8. Pedestrian motion comparison between real-world and generated video. Top row: real footage at $t = 3$ s, $t = 4$ s, $t = 5$ s. Bottom row: corresponding generated frames. The pedestrian moves at an unrealistically slow pace and remains near the right side of the road throughout the sequence.

The second failure mode involves depth drift and geometric instability. Moderate 3D consistency scores correspond to occasional fluctuations in object scale or perspective during motion. Figure 9 illustrates this phenomenon: a female pedestrian walks toward the ego vehicle, initially appearing at a plausible scale in the distance. However, her figure grows at an abnormally rapid rate as she approaches, and by the final frame, she towers above the SUV parked on the right side of the scene, which is a clear violation of real-world proportions. While each individual frame is geometrically coherent in isolation, the temporal progression of scale lacks consistency with the estimated depth. Such artefacts are typically subtle in shorter clips but become more pronounced in longer sequences, where compounding depth estimation errors gradually distort object proportions.

Finally, object-level persistence inconsistencies occur in a subset of videos. Minor scene elements, such as traffic signs or distant vehicles, sometimes exhibit texture drift or partial disappearance across frames (e.g., Figure 6), which contributes to moderate photometric consistency scores and lower object-level human ratings.



Figure 9. Three frames from a generated video clip illustrating depth drift: (a) at $t = 1$, the pedestrian approaching the ego vehicle appears at a plausible distance; (b) at $t = 3$, anomalous scale growth becomes apparent as the pedestrian enlarges rapidly; (c) at $t = 5$, the pedestrian towers above the reference SUV visible on the right side.

8. Automatic Metrics Versus Human Perception

A notable observation is the discrepancy between automatic and human evaluations of motion quality. While motion smoothness achieves a near-perfect score (mean score: 0.99), human ratings for motion alignment are substantially lower. This indicates that low-level temporal stability metrics are insufficient to capture higher-level behavioural correctness and intent consistency. A similar ceiling effect is observed in other metrics, such as content alignment and human anatomy, which consistently achieve near-saturated scores with minimal variance. Although these results suggest strong overall performance, they provide limited discriminative power for ranking samples based on motion realism or behavioural plausibility.

In addition, metrics intended to capture visual consistency present their own limitations. Photometric consistency exhibits large dispersion across samples, yet this variability does not align with human perception, which rates environmental realism and scene fidelity highly. This mismatch reflects the inherent unsuitability of photometric metrics in dynamic driving scenarios, where illumination changes are natural and expected rather than indicative of generation errors. As a result, such metrics may over-penalise realistic variations while failing to reflect perceptual quality.

These findings reveal a broader limitation of metric-based evaluation: existing quantitative measures are effective at capturing low-level visual fidelity and coarse semantic alignment but remain insufficient for assessing medium to high-level properties, such as behavioural plausibility, motion grounding, and contextual coherence. This limitation is particularly critical in safety-sensitive domains like autonomous driving, where higher-level semantic correctness can be critical. Consequently, these results highlight the necessity of combining quantitative generative benchmarks with human-centred evaluation, where human judgement provides essential insight into real-world plausibility and behavioural validity that automated metrics fail to capture.

The evaluation demonstrates that SynPeDAS provides high-quality, temporally stable, and prompt-aligned synthetic pedestrian videos that can be used for data augmentation. These results suggest that, when appropriate quality control is applied, the generated data may be more suitable for appearance-oriented exploratory augmentation than for tasks requiring physically grounded long-term motion reasoning. Systematic evaluation on tracking, trajectory prediction, intent prediction, and risk assessment is left for future work.

Overall, SynPeDAS represents a practical step toward scalable synthetic data generation for pedestrian perception in autonomous driving. At the same time, the results emphasise the need for improved motion grounding and geometry-aware constraints in future generative world models to meet the strict reliability requirements of safety-critical systems.

9. Conclusions

This paper investigated whether current world foundation models can generate pedestrian-focused driving videos that are suitable for autonomous driving research, and whether commonly used evaluation metrics can reveal the safety-critical failures that arise in such generated scenarios. To support this investigation, we introduced SynPeDAS, an open research framework that integrates a reproducible generation pipeline, an automated evaluation suite, a structured human perceptual assessment, and downstream demonstration code. By instantiating the framework with NVIDIA Cosmos Predict 2.5 and anchoring generation in real pedestrian–vehicle interaction scenes, this study provides a systematic case study of both the opportunities and limitations of world model-based synthetic data generation for autonomous driving.

The results show that the generated videos achieve strong visual fidelity, high prompt adherence, and good short-term temporal smoothness. These strengths suggest that world model-generated data may be useful for controlled data exploration and appearance-oriented augmentation when appropriate quality control is applied. Nevertheless, our analysis also identifies recurring failure modes, including dynamic misalignment, depth drift and geometric instability, and object persistence issues. These limitations are especially problematic for tasks that depend on physically grounded motion and behaviourally plausible interactions, such as pedestrian trajectory modelling, intent prediction, and risk assessment. Therefore, synthetic videos generated by current world foundation models should not be treated as a reliable replacement for real-world driving data in safety-critical applications.

Our findings further demonstrate that existing automatic evaluation metrics are insufficient for judging the reliability of generated driving videos. Many commonly used metrics exhibit ceiling effects and fail to capture subtle but safety-critical errors in motion and behaviour. Many metrics overemphasise visual fidelity or coarse semantic similarity while failing to capture safety-critical errors in motion, geometry, and behaviour. The weak alignment between metric scores and human judgement indicates that safety-oriented evaluation requires more than generic video generation benchmarks. In particular, future evaluation protocols should explicitly assess motion grounding, geometric consistency, behavioural plausibility, and downstream task relevance.

SynPeDAS shows that world foundation models offer a promising but still limited route toward scalable synthetic data generation for pedestrian-centric autonomous driving research. The framework provides a reproducible platform for analysing generated driving videos, identifying model failure modes, and studying the gap between automatic metrics and human perception. Future work should focus on geometry-aware and physics-informed generation constraints, domain-specific safety metrics, broader rare-scenario coverage, and systematic downstream experiments to determine when synthetic data improves, degrades, or biases autonomous driving perception models.

Author Contributions: Conceptualisation, Q.L.; Data Curation and Formal Analysis, C.Z.; Funding Acquisition, Q.L.; Methodology, Q.L., C.Z. and S.A.; Supervision, Q.L., O.H. and V.P.; Writing, C.Z. and S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the EPSRC and the AI Security Institute under Grant Reference UKRI855. The first author gratefully acknowledges the university scholarship provided by Coventry University.

Data Availability Statement: Our framework can be accessed through this GitHub repository link: <https://github.com/levilovearch/SynPeDAS> (accessed on 6 July 2016).

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

Appendix A. Filtering Conditions for Pedestrian Crossing Events

We construct a filtered subset of the JAAD dataset using the following conditions:

1. **Minimum video duration:** Only video clips with a total duration longer than 5 s are considered.
2. **Selection under multiple pedestrians:** In cases where multiple pedestrians are present and crossing the road within a clip, we select the temporal segment with the highest pedestrian density (i.e., the time window where the maximum number of pedestrians are simultaneously engaged in crossing behaviour).
3. **Temporal padding for short crossing events:** For each selected crossing event:
 - If the total duration of the crossing behaviour is less than 5 s, we extend the clip to reach a minimum length of 5 s.
 - The extension is performed by first including frames preceding the onset of the crossing action.
 - If the preceding frames are insufficient to reach 5 s, additional frames are appended after the crossing event until the minimum duration requirement is satisfied.

Appendix B. Instructions for VLM to Generate Caption for Video Clips

1. Follow the ****6-Layer Model (6LM)**** to structure each scene conceptually:
 - * ****Layer 1 (Road):**** Describe road category, markings, or surface details.
 - * ****Layer 2 (Roadside Structures):**** Include buildings, sidewalks, or urban fixtures.
 - * ****Layer 3 (Temporal Modifications):**** Reflect dynamic or transitional elements like motion, time of day, or changes in lighting.
 - * ****Layer 4 (Objects):**** Focus on vehicles, pedestrians, cyclists, or other moving and static elements.
 - * ****Layer 5 (Daytime & Weather):**** Accurately represent illumination, shadows, and environmental conditions.
 - * ****Layer 6 (Communication):**** Integrate visible signaling or informational sources such as traffic lights, digital signs, or road markings.
2. Describe only the ****first 5 seconds**** of the scene. Capture a specific, self-contained moment without referencing time stamps or future actions.
3. ****Emphasize physical realism****:
 - * Ensure all objects behave according to real-world physics and natural laws.
 - * Maintain logical spatial relationships-objects should occupy believable positions and scales.
 - * Avoid surreal or impossible geometry unless explicitly requested.
4. ****Focus on concrete sensory details****:
 - * Specify materials (e.g., wet asphalt, metal railing, glass panel).
 - * Note lighting properties (e.g., diffused sunlight, reflected highlights).
 - * Mention textures, reflections, shadows, and occlusion effects.
5. ****Incorporate photographic and cinematographic language****:
 - * Refer to composition (e.g., low-angle shot, wide framing, centered subject).
 - * Describe lighting setups (e.g., natural backlight, soft fill).

- * Mention camera attributes when relevant (e.g., shallow depth of field, steady wide-angle lens, medium shot).
6. Keep the **tone natural and grounded**:
- * Do not use bullet points or lists.
 - * Write as continuous descriptive prose, in fewer than **250 words**.
 - * Do not include negative prompts or disallowed content.
 - * Do not mention specific vehicle brands or models.
7. Maintain **logical consistency and narrative flow**:
- * Ensure that all elements belong to the same spatial and lighting environment.
 - * The scene should progress smoothly and remain temporally coherent.
 - * Avoid contradictions such as mismatched lighting directions or impossible reflections.

Appendix C. Prompt Samples

Appendix C.1. Prompts Used to Generate the Video Shown in Figure 4

A wide-angle, dashcam perspective captures a sunlit urban street, its cracked asphalt surface marked with subtle damp patches. On the left, a row of mature, leafy green trees shades a narrow sidewalk, running alongside a multi-story brick building and several tall utility poles. Across the street, another brick structure is similarly verdant, with a dense canopy of trees casting deep shadows onto the pavement. A pristine white sports car is parked curbside on the right, its glossy finish reflecting the bright midday sun. As the scene unfolds, several men, casually dressed, walk from the right foreground, their figures casting distinct, elongated shadows on the street as they cross the vehicle's path. They proceed towards the left, eventually turning into a narrow alleyway that extends between more brick buildings and a line of tall, slender trees, revealing other parked cars in the distance. The environment is bathed in natural daylight, highlighting the vibrant green foliage against the warm brick tones under a clear blue sky.

Appendix C.2. Prompts Used to Generate the Video Shown in Figure 7

From a steady, wide-angle dashboard perspective, an urban street unfolds under a bright, diffused sky. A multi-lane road stretches forward, marked by the presence of a central divider on the left, which hosts a striking black metal sculpture depicting abstract human figures in a dynamic, forward-leaning stance. This art installation stands before modern glass skyscrapers that gleam with reflections of the overcast sky, their sleek facades soaring skyward. On the right, a sequence of brick and concrete buildings with large storefronts, including a Best Buy, lines the bustling sidewalk. Traffic ahead appears largely stationary, consisting of various vehicles, including a dark-coloured sedan and a silver SUV in the adjacent lane. A pedestrian, dressed in dark attire, begins to cross the street from the left sidewalk, moving purposefully through the stopped traffic lanes toward the right side of the frame, their silhouette momentarily cutting across the urban panorama.

References

1. Ding, J.; Zhang, Y.; Shang, Y.; Zhang, Y.; Zong, Z.; Feng, J.; Yuan, Y.; Su, H.; Li, N.; Sukiennik, N.; et al. Understanding world or predicting future? A comprehensive survey of world models. *ACM Comput. Surv.* **2025**, *58*, 57. [\[CrossRef\]](#)
2. Jiang, C.M.; Masotto, X.; Sun, B. The Waymo World Model: A New Frontier for Autonomous Driving Simulation. *Waypoint: The Official Waymo Blog*, 6 February 2026. Available online: <https://waymo.com/blog/2026/02/the-waymo-world-model-a-new-frontier-for-autonomous-driving-simulation/> (accessed on 6 July 2026).

3. Awais, M.; Naseer, M.; Khan, S.; Anwer, R.M.; Cholakkal, H.; Shah, M.; Yang, M.H.; Khan, F.S. Foundation models defining a new era in vision: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 2245–2264. [CrossRef] [PubMed]
4. Agarwal, N.; Ali, A.; Bala, M.; Balaji, Y.; Barker, E.; Cai, T.; Chattopadhyay, P.; Chen, Y.; Cui, Y.; Ding, Y.; et al. Cosmos world foundation model platform for physical ai. *arXiv* **2025**, arXiv:2501.03575.
5. Rasouli, A.; Kotseruba, I.; Tsotsos, J.K. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 206–213.
6. Rasouli, A.; Kotseruba, I.; Kunic, T.; Tsotsos, J.K. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6262–6271.
7. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
8. Dollár, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: A benchmark. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Miami, FL, USA, 2009; pp. 304–311.
9. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Koltun, V. CARLA: An open urban driving simulator. In *Proceedings of the Conference on Robot Learning*; PMLR: Mountain View, CA, USA, 2017; pp. 1–16.
10. Tarel, J.P.; Hautiere, N.; Cord, A.; Gruyer, D.; Halmaoui, H. Improved visibility of road scene images under heterogeneous fog. In *Proceedings of the 2010 IEEE Intelligent Vehicles Symposium*; IEEE: La Jolla, CA, USA, 2010; pp. 478–485.
11. Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; Lopez, A.M. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3234–3243. [CrossRef]
12. Gaidon, A.; Wang, Q.; Cabon, Y.; Vig, E. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4340–4349.
13. Wrenninge, M.; Unger, J. Synscapes: A Photorealistic Synthetic Dataset for Street Scene Parsing. *arXiv* **2018**, arXiv:1810.08705.
14. Apollo Auto. Apollo Synthetic Dataset. Available online: <https://developer.apollo.auto/synthetic.html> (accessed on 26 June 2026).
15. 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.
16. Park, T.; Liu, M.Y.; Wang, T.C.; Zhu, J.Y. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2337–2346.
17. Tang, H.; Xu, D.; Sebe, N.; Wang, Y.; Corso, J.J.; Yan, Y. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2417–2426.
18. Lee, C.H.; Liu, Z.; Wu, L.; Luo, P. Maskgan: Towards diverse and interactive facial image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5549–5558.
19. Wen, Y.; Zhao, Y.; Liu, Y.; Jia, F.; Wang, Y.; Luo, C.; Zhang, C.; Wang, T.; Sun, X.; Zhang, X. Panacea: Panoramic and controllable video generation for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 6902–6912.
20. Song, Z.; He, Z.; Li, X.; Ma, Q.; Ming, R.; Mao, Z.; Pei, H.; Peng, L.; Hu, J.; Yao, D.; et al. Synthetic datasets for autonomous driving: A survey. *IEEE Trans. Intell. Veh.* **2023**, *9*, 1847–1864. [CrossRef]
21. Murray, R.M.; Li, Z.; Sastry, S.S. *A mathematical Introduction to Robotic Manipulation*; CRC Press: Boca Raton, FL, USA, 2017.
22. Lennart, L. *System Identification: Theory for the User*; PTR Prentice Hall: Upper Saddle River, NJ, USA, 1999; Volume 28, p. 540.
23. Ha, D.; Schmidhuber, J. World models. *arXiv* **2018**, *2*, 440.
24. Hafner, D.; Lillicrap, T.; Ba, J.; Norouzi, M. Dream to Control: Learning Behaviors by Latent Imagination. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26 April–1 May 2020.
25. Hafner, D.; Lillicrap, T.P.; Norouzi, M.; Ba, J. Mastering Atari with Discrete World Models. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
26. Hafner, D.; Pasukonis, J.; Ba, J.; Lillicrap, T. Mastering diverse domains through world models. *arXiv* **2023**, arXiv:2301.04104.
27. Hansen, N.; Su, H.; Wang, X. Td-mpc2: Scalable, robust world models for continuous control. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024; Volume 2024, pp. 47376–47405.
28. Bruce, J.; Dennis, M.D.; Edwards, A.; Parker-Holder, J.; Shi, Y.; Hughes, E.; Lai, M.; Mavalankar, A.; Steigerwald, R.; Apps, C.; et al. Genie: Generative interactive environments. In Proceedings of the Forty-First International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024.
29. Micheli, V.; Alonso, E.; Fleuret, F. Transformers are Sample-Efficient World Models. In Proceedings of the the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
30. Alonso, E.; Jelley, A.; Micheli, V.; Kanervisto, A.; Storkey, A.; Pearce, T.; Fleuret, F. Diffusion for world modeling: Visual details matter in atari. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 58757–58791. [CrossRef]

31. Hua, P.; Liu, M.; Macaluso, A.; Lin, Y.; Zhang, W.; Xu, H.; Wang, L. Gensim2: Scaling robot data generation with multi-modal and reasoning llms. In Proceedings of the 8th Conference on Robot Learning, Munich, Germany, 6–9 November 2024; pp. 5030–5066.
32. Nasiriany, S.; Maddukuri, A.; Zhang, L.; Parikh, A.; Lo, A.; Joshi, A.; Mandlekar, A.; Zhu, Y. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In Proceedings of the Robotics: Science and Systems, Delft, The Netherlands, 15–19 July 2024.
33. Assran, M.; Bardes, A.; Fan, D.; Garrido, Q.; Howes, R.; Muckley, M.; Rizvi, A.; Roberts, C.; Sinha, K.; Zholus, A.; et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv* **2025**, arXiv:2506.09985.
34. Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 7514–7528.
35. Duan, H.; Yu, H.X.; Chen, S.; Fei-Fei, L.; Wu, J. Worldscore: A unified evaluation benchmark for world generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 27713–27724.
36. Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. Vbench: Comprehensive benchmark suite for video generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 21807–21818.
37. Zheng, D.; Huang, Z.; Liu, H.; Zou, K.; He, Y.; Zhang, F.; Gu, L.; Zhang, Y.; He, J.; Zheng, W.S.; et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv* **2025**, arXiv:2503.21755.
38. Liu, Y.; Zhou, H.; Guo, Z.; Shareghi, E.; Vulić, I.; Korhonen, A.; Collier, N. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In Proceedings of the First Conference on Language Modeling, Philadelphia, PA, USA, 7–9 October 2024.
39. Li, D.; Fang, Y.; Chen, Y.; Yang, S.; Cao, S.; Wong, J.; Luo, M.; Wang, X.; Yin, H.; Gonzalez, J.E.; et al. WorldModelBench: Judging Video Generation Models as World Models. *Adv. Neural Inf. Process. Syst.* **2026**, *38*. [[CrossRef](#)]
40. Google. Gemini 2.5 Flash. Available online: <https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash> (accessed on 6 July 2026).
41. Scholtes, M.; Westhofen, L.; Turner, L.R.; Lotto, K.; Schuldes, M.; Weber, H.; Wagener, N.; Neurohr, C.; Bollmann, M.H.; Körtke, F.; et al. 6-layer model for a structured description and categorization of urban traffic and environment. *IEEE Access* **2021**, *9*, 59131–59147. [[CrossRef](#)]
42. NVIDIA. Cosmos-Cookbook: Post-Training Scripts and Samples for NVIDIA Cosmos Ecosystem. Available online: <https://github.com/nvidia-cosmos/cosmos-cookbook> (accessed on 6 July 2026).
43. Lee, Y.; Park, I.; Kang, M. *FLEUR: An Explainable Reference-Free Evaluation Metric for Image Captioning Using a Large Multimodal Model*; Association for Computational Linguistics: Bangkok, Thailand, 2024.
44. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved Baselines with Visual Instruction Tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 26296–26306.
45. Teed, Z.; Deng, J. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 16558–16569.
46. 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. In Proceedings of the International Conference on Machine Learning. PmlR, Online, 18–24 July 2021; pp. 8748–8763.
47. Li, Z.; Zhu, Z.L.; Han, L.H.; Hou, Q.; Guo, C.L.; Cheng, M.M. Amt: All-pairs multi-field transforms for efficient frame interpolation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 9801–9810.
48. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6627–6638.
49. Jayasumana, S.; Ramalingam, S.; Veit, A.; Glasner, D.; Chakrabarti, A.; Kumar, S. Rethinking FID: Towards a Better Evaluation Metric for Image Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024.
50. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.Y.; Girshick, R. Detectron2. 2019. Available online: <https://github.com/facebookresearch/detectron2> (accessed on 6 July 2026).
51. Jocher, G.; Qiu, J. Ultralytics YOLO11, Version 11.0.0. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 6 July 2026).

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.