WorldLens: Full-Spectrum Evaluations of Driving World Models in Real World

Ao Liang 🚕, **Lingdong Kong** 🚕🚙, **Tianyi Yan** 🚕, **Hongsi Liu** 🚕, **Wesley Yang** 🚕, **Ziqi Huang**, **Wei Yin**, **Jialong Zuo**, **Yixuan Hu**, **Dekai Zhu**, **Dongyue Lu**, **Youquan Liu**, **Guangfeng Jiang**, **Linfeng Li**, **Xiangtai Li**, **Long Zhuo**, **Lai Xing Ng**, **Benoit R. Cottereau**, **Changxin Gao**, **Liang Pan**, **Wei Tsang Ooi**, **Ziwei Liu** 🛻

WorldBench Team

🚕 **Equal Contributions** 🚙 **Project Lead** 🛻 **Corresponding Author**

Generative world models are reshaping embodied AI, enabling agents to synthesize realistic 4D driving environments that look convincing but often fail physically or behaviorally. Despite rapid progress, the field still lacks a unified way to assess whether generated worlds preserve geometry, obey physics, or support reliable control. We introduce **WorldLens**, a full-spectrum benchmark evaluating how well a model builds, understands, and behaves within its generated world. It spans five aspects – Generation, Reconstruction, Action-Following, Downstream Task, and Human Preference – jointly covering visual realism, geometric consistency, physical plausibility, and functional reliability. Across these dimensions, no existing world model excels universally: those with strong textures often violate physics, while geometry-stable ones lack behavioral fidelity. To align objective metrics with human judgment, we further construct **WorldLens-26K**, a large-scale dataset of human-annotated videos with numerical scores and textual rationales, and develop **WorldLens-Agent**, an evaluation model distilled from these annotations to enable scalable, explainable scoring. Together, the benchmark, dataset, and agent form a unified ecosystem for measuring world fidelity – standardizing how future models are judged not only by how real they look, but by how real they behave.

Project Page: https://worldbench.github.io/worldlens

GitHub Repo: https://github.com/worldbench/WorldLens

HuggingFace Leaderboard: https://huggingface.co/spaces/worldbench/WorldLens

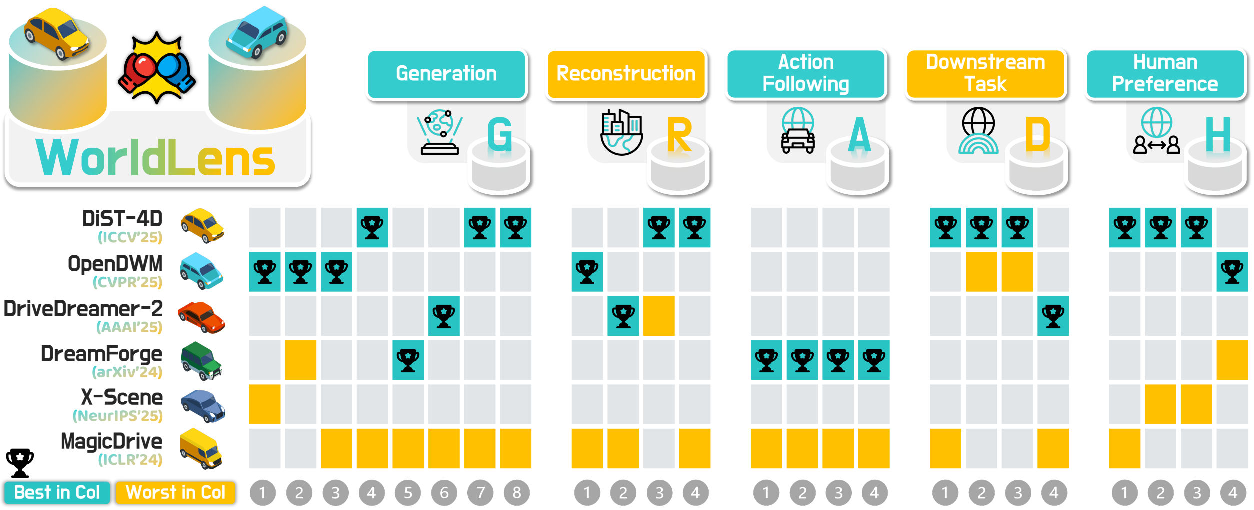
HuggingFace Dataset: https://huggingface.co/datasets/worldbench

1 Introduction

Generative world models have transformed embodied AI and simulation [3, 77, 82, 86, 136]. From text-driven 4D synthesis to controllable driving environments [72, 89, 103, 104, 121, 139], modern systems can produce dash-cam–like sequences with striking visual realism. However, evaluation has not kept pace: the field lacks a standardized way to measure whether generated worlds preserve geometry, respect physics, and support reliable decision-making [53, 78, 107].

Most widely used metrics emphasize frame quality and aesthetics [1, 41, 42, 49], but reveal little about physical causality, multi-view geometry, or functional reliability under control [5, 29, 120, 124, 131, 143, 146]. This gap, which is well documented across recent surveys [53, 68, 107, 141], has created fragmented progress and incomparable results. While structured maturity scales, *e.g.*, SAE Levels of Driving Automation$^{\mathrm{TM}}$, have clarified autonomy benchmarking [73], an analogous, practice-ready protocol for **evaluating driving world models** has remained elusive.

To bridge the gap, we build **WorldLens**, a full-spectrum benchmark that evaluates how well a world model



	Generation								Reconstruction				Action Following				Downstream Task				Human Preference			
	1	2	3	4	5	6	7	8	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
DiST-4D (ICCV'25)				🏆			🏆	🏆			🏆	🏆					🏆	🏆	🏆		🏆	🏆	🏆	
OpenDWM (CVPR'25)	🏆	🏆	🏆						🏆									🟨	🟨					🏆
DriveDreamer-2 (AAAI'25)						🏆				🏆	🟨									🏆				
DreamForge (arXiv'24)		🟨			🏆								🏆	🏆	🏆	🏆								🟨
X-Scene (NeurIPS'25)	🟨																					🟨	🟨	
MagicDrive (ICLR'24)			🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨		🟨	🟨	🟨	🟨	🟨	🟨			🟨	🟨			

🏆 Best in Col 🟨 Worst in Col

Figure 1 Is your driving world model an all-around player? This work presents **WorldLens**, a unified benchmark encompassing evaluations on [1]**Generation**, [2]**Reconstruction**, [3]**Action-Following**, [4]**Downstream Task**, and [5]**Human Preference**, across a total of **24 dimensions** spanning visual realism, geometric consistency, functional reliability, and perceptual alignment. We observe *no single model dominates across all axes*, highlighting the need for balanced progress toward physically and behaviorally realistic world modeling.

builds, *understands*, and *behaves* within its generated world. As shown in Figure 1, no existing model excels universally; some achieve strong texture realism but violate physics, while others preserve geometry yet fail behaviorally.

To reveal these world modeling trade-offs, we decompose evaluation into **five complementary aspects**:

- [1]**Generation** — measuring whether a model can synthesize *visually realistic*, *temporally stable*, and *semantically consistent* scenes [27, 30, 89]. Even state-of-the-art models that achieve low perceptual error (*e.g.*, LPIPS, FVD) often suffer from view flickering or motion instability, revealing the limits of current diffusion-based architectures.
- [2]**Reconstruction** — probing whether generated videos can be reprojected into a *coherent 4D scene* using differentiable rendering [17, 50]. Models that appear sharp in 2D frequently collapse when reconstructed, producing geometric "floaters": a gap that exposes how temporal coherence remains weakly coupled in most pipelines.
- [3]**Action-Following** — testing if a pre-trained action planner [38, 45] can *operate safely* inside the generated world. High *open-loop* realism does not guarantee safe *closed-loop* control; almost all existing world models trigger collisions or off-road drifts, underscoring that photometric realism alone cannot yield functional fidelity.
- [4]**Downstream Task** — evaluating whether the *synthetic data* support downstream perception models trained on real-world datasets [9, 66, 96]. Even visually appealing worlds may degrade detection or segmentation accuracy by 30–50%, highlighting that alignment to task distributions, not just image quality, is vital for practical usability.
- [5]**Human Preference** — capturing subjective scores such as *world realism*, *physical plausibility*, and *behavioral safety* through large-scale human annotations. Our study reveals that models with strong geometric consistency are generally rated as more "real", confirming that perceptual fidelity is inseparable from structural coherence.

To bridge algorithmic metrics with human perception, we curate **WorldLens-26K**, a large-scale dataset of human-annotated videos covering perceptual, physical, and safety-related dimensions. Each sample contains quantitative scores and textual explanations, capturing how evaluators reason about realism, physical plausibility, and behavioral safety. By pairing human judgments with structured rationales, we aim to transform subjective evaluation into learnable supervision, enabling perception-aligned and interpretable assessment of generative world models.

Leveraging the above, we develop **WorldLens-Agent**, a feedback-aligned auto-evaluator distilled from human preferences. This agent can predict perceptual and physical scores while generating natural-language explanations consistent with human reasoning. It generalizes well to *unseen* models and enables scalable auto-evaluation of generative worlds without repeated manual labeling.

Together, our benchmark, dataset, and evaluation agent form **a unified ecosystem** that bridges objective measurement and human interpretation. *We will release the toolkit, dataset, and model to foster standardized, explainable, and human-aligned evaluation* – guiding future world models not only to *"look"* real, but to *"behave"* reasonably.

2 Related Work

Video Generation. Recent advances have driven rapid progress in generation [4, 55, 68, 106, 119, 136]. Text-to-image [6, 85, 87, 114] laid the foundation for high-fidelity synthesis from text, later extended to the temporal domain through text-to-video (T2V) systems [8, 26, 36, 43, 54, 74, 82, 90, 91, 97–100, 132]. Building on these foundations, domain-specific methods [27, 32, 56, 62, 67, 94, 105, 116, 145] achieved impressive realism using motion-aware conditioning [7, 39, 44, 46, 89, 108, 115]. Despite strong perceptual quality, they remain largely *appearance-driven*: they generate visually coherent sequences but lack explicit geometry, physics, or causal control [5, 14, 120, 121, 130]. Without structured world states or dynamics, they cannot model how scenes behave or respond to actions [1, 20, 125], and metrics focused only on visual fidelity reveal little about whether a model truly understands the world it depicts.

3D & 4D World Modeling. Recent studies move beyond frame-based generation to build *world models* that encode 3D, dynamics, and control-aware representations [53]. WonderWorld [135], GAIA-1/2 [37, 86], Genie-3 [3], and related efforts [29, 31, 124, 133] learn physics-grounded latent states representing geometry, occupancy, and motion for conditioned predictions [5, 16, 52, 59, 60, 65, 70, 129, 147]. DriveArena [127] and NAVSIM [11, 20] integrate perception, planning, and control into generative pipelines that support closed-loop simulation [10, 23, 123]. Yet, existing paradigms remain limited to perception metrics or qualitative results [53]. A unified standard for measuring geometry, 4D consistency, and agent behaviors is missing.

Evaluation. Generative video evaluation has evolved from simple frame-based scores to multi-dimensional benchmarks [57, 61, 63, 102, 137]. Early metrics measure distributional similarity and perceptual alignment [35, 49, 79, 88, 101], while frameworks, *e.g.*, VBench [41, 42], EvalCrafter [64], and T2V-CompBench [93], extend assessment to motion and temporal consistency. More recent efforts, WorldScore [25] and VideoWorld [84], move toward "world-model" evaluation using physics-inspired composite scores, yet they remain confined to 2D video settings emphasizing appearance over embodiment [47]. In driving, some existing benchmarks [10, 23, 118] evaluate agents rather than the worlds they inhabit. **WorldLens** introduces 4D reconstruction, action-following, and human preference alignment to jointly assess spatial fidelity, behavioral consistency, and physical realism, establishing the **first** benchmark that measures both the appearance and behavior.

3 WorldLens: A Full-Spectrum Benchmark

Generative world models must go beyond visual realism to achieve geometric consistency, physical plausibility, and functional reliability. We introduce **WorldLens**, a unified benchmark that evaluates these capabilities across five complementary aspects – from **low-level** appearance fidelity to **high-level** behavioral realism. As shown in Figure 2, each aspect is decomposed into fine-grained, interpretable dimensions, forming a comprehensive framework that bridges human perception, physical reasoning, and downstream utility.

3.1 Generation

This aspect decomposes the overall **generation quality** into eight complementary dimensions that assess appearance fidelity, temporal stability, geometric correctness, and semantic smoothness. Together, these

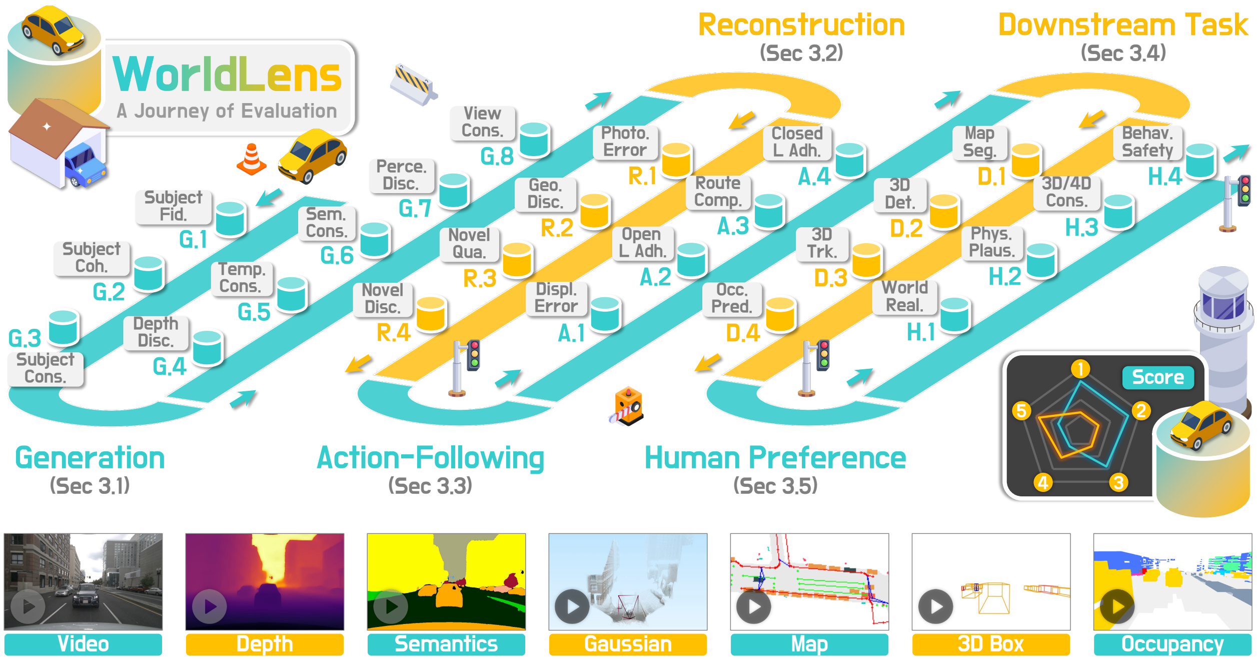


Figure 2 The WorldLens evaluation framework unifies five complementary aspects – ^{1}**Generation**, ^{2}**Reconstruction**, ^{3}**Action-Following**, ^{4}**Downstream Task**, and ^{5}**Human Preference** – to assess visual, geometric, functional, and perceptual fidelity of generative world models. Each aspect is decomposed into interpretable dimensions driven by measurable signals such as segmentation, depth, 4D reconstruction, and behavioral simulation, enabling comprehensive and physically grounded evaluations across the full spectrum of world modeling.

dimensions quantify how "faithfully" a model constructs visually and perceptually consistent driving scenes across time and viewpoints.

G.1 **Subject Fidelity** measures the perceptual realism and semantic correctness of individual object instances, *e.g.*, vehicles and pedestrians. For each generated frame, object regions are cropped using bounding boxes and evaluated with class-specific binary classifiers trained on real data [24]. A high confidence indicates that the generated object visually aligns with its real-world category. This dimension captures localized realism and complements global metrics by focusing on instance-level fidelity.

G.2 **Subject Coherence** assesses the temporal stability of each object's identity throughout a generated sequence. Using known track IDs, we extract visual embeddings for each instance from a pretrained ReID model [34, 148] and compute similarity across frames. A high coherence score indicates that objects maintain consistent shape, color, and texture through motion, reflecting stable appearance and identity preservation over time.

G.3 **Subject Consistency** evaluates fine-grained temporal stability of object-level semantics and geometry. It uses DINO [12] features to capture texture and spatial details across frames, ensuring that subjects preserve their structure and semantic meaning without flickering or deformation. High consistency reflects the reliable temporal evolution of objects and smooth changes in appearance under motion.

G.4 **Depth Discrepancy** measures the smoothness of depth variations across time, capturing geometric coherence. We estimate monocular depth for each frame using Depth Anything V2 [126], coded with colors, and extract corresponding embeddings through DINO v2 [75]. The average feature distance between consecutive frames quantifies the continuity of depth perception. Lower discrepancy indicates geometrically stable and physically plausible scene motion.

G.5 **Temporal Consistency** quantifies global frame-to-frame smoothness in a learned appearance space. Each frame is embedded with the CLIP visual encoder [79], and temporal stability is derived from adjacent-frame similarity, jitter suppression, and motion-rate alignment with real videos. High consistency corresponds to temporally stable dynamics without abrupt or unnatural transitions.

G.6 **Semantic Consistency** evaluates whether the semantic layout of generated scenes evolves smoothly across time. We employ a pretrained SegFormer [117] to predict frame-wise masks and compute stability across labels, regions, and class distributions. This dimension ensures that generated videos maintain coherent object semantics and scene structures without flickering boundaries or label inconsistencies.

G.7 **Perceptual Discrepancy** quantifies the overall perceptual gap between real and generated videos. We extract spatiotemporal embeddings from a pretrained I3D [13] and compute the Fréchet Video Distance [101] between real and generated distributions. A lower discrepancy indicates closer alignment in appearance and motion statistics, reflecting perceptual realism and temporal coherence.

G.8 **Cross-View Consistency** measures the geometric and photometric alignment between overlapping regions of adjacent camera views. Using LoFTR [92], we detect feature correspondences between synchronized camera pairs and aggregate their confidence to assess spatial coherence. High consistency indicates better structural alignment and visual continuity across multiple perspectives, ensuring 3D-consistent generation for multi-camera driving systems.

3.2 Reconstruction

This aspect decomposes the overall **reconstructability** into how well a coherent 4D scene can be recovered from generated videos. Each sequence is lifted into a Gaussian Field and re-rendered under both original and **novel camera trajectories**, testing spatial interpolation, parallax, and view generalization across representative novel-view paths.

R.1 **Photometric Error** measures how accurately reconstructed scenes reproduce their input frames. Each generated video is reconstructed into a differentiable 4D representation [17, 50], and re-rendered at original camera poses. Pixel-level similarities, *i.e.*, LPIPS, PSNR, and SSIM, are computed, reflecting stability of appearance and lighting across time. Lower discrepancy indicates more consistent photometric properties for faithful 4D reconstruction.

R.2 **Geometric Discrepancy** assesses how well the reconstructed geometry from generated videos aligns with real-world structure. Using identical camera poses, we reconstruct 4D scenes from both generated and ground-truth sequences and compare their rendered depth maps. The Absolute Relative Error (AbsRel) and other related metrics are computed within regions selected by Grounded-SAM 2 [80, 81]. Lower values indicate more plausible depth and consistent surface geometry with real scenes.

R.3 **Novel-View Quality** evaluates the perceptual realism of re-rendered frames from unseen camera trajectories. Each reconstructed scene is rendered along novel paths using the same differentiable framework, and visual quality is scored by MUSIQ [49]. A higher score indicates that novel views remain sharp, artifact-free, and visually coherent, demonstrating that the model preserves appearance and illumination consistency beyond training viewpoints.

R.4 **Novel-View Discrepancy** quantifies the perceptual gap between novel-view renderings from generated and real reconstructions. Both are rendered under identical camera trajectories, and their distance is measured via FVD [101] on I3D features [13]. Lower discrepancy indicates better generalization to unseen viewpoints, maintaining coherent geometry, appearance, and temporal dynamics in 4D space.

3.3 Action-Following

This aspect evaluates how well the generated worlds support **plausible driving decisions** when interpreted by pretrained planners, examining whether synthesized scenes provide realistic visual and motion cues that yield real-world-consistent actions. All evaluations are conducted in a generative simulator using custom-designed routes derived from real-world maps against standard benchmarks [9, 10].

A.1 **Displacement Error** measures functional consistency between the trajectories predicted from generated and real videos. Using a pretrained end-to-end planner, UniAD [38] or VAD [45], both sequences are used to predict future waypoints, and their mean L2 distance is computed. Lower displacement error indicates that the generated scenes preserve motion cues required for accurate trajectory forecasting.

A.2 **Open-Loop Adherence** evaluates how well a pretrained policy performs when operating on generated videos in a generative simulator. In open-loop mode, the policy's predictions are used purely for evaluation and do not affect the simulated ego-vehicle motion. Following NAVSIM [20], we adopt the Predictive Driver Model Score (PDMS), which aggregates safety, progress, and comfort sub-scores computed over a short simulation horizon. Higher PDMS indicates that the policy exhibits realistic, stable, and safe driving behaviors even when guided solely by generated visual input, reflecting reliable short-term functional realism.

A.3 **Route Completion** measures long-horizon navigation stability in closed-loop simulation. It computes the percentage of a predefined route completed before termination due to collision, off-road drift, or timeout. Higher route completion rates signify that generated environments support continuous, physically consistent control over extended trajectories, enabling sustained goal-directed motion.

A.4 **Closed-Loop Adherence** integrates both motion quality and task success into a single metric, the Arena Driving Score (ADS) [127]. In the closed-loop mode, the planning decisions of the driving agent directly control the ego's actions, thereby influencing the simulated environment. ADS multiplies the PDMS and Route Completion scores, rewarding agents that are both safe and successful. A high ADS implies that the planner achieves realistic, collision-free navigation while completing the route effectively, confirming that generated worlds not only *look* real but also *drive* real within an autonomous-control loop.

3.4 Downstream Task

This aspect evaluates the **downstream utility** of generated videos by measuring how well 3D perception models pretrained on real data perform when applied to synthetic content. Performance degradation across tasks reflects the realism, fidelity, and transferability of generated scenes.

D.1 **Map Segmentation** evaluates whether generated data contain sufficient spatial and semantic cues for top-down BEV mapping. A pretrained BEV map segmentation network [58, 66] predicts semantic maps for each frame, and performance is measured by mean Intersection-over-Union (mIoU). Higher mIoU indicates better structural layout and semantics conducive to accurate BEV reconstruction.

D.2 **3D Object Detection** tests whether generated data retains geometric cues essential for perceiving traffic participants. A pretrained BEVFusion [66] is applied to generated frames, and performance is reported using nuScenes Detection Score (NDS) [9]. Higher values indicate that generated scenes support more accurate 3D object localization.

D.3 **3D Object Tracking** measures motion consistency and identity information preservation of generated videos across time. A pretrained 3D tracker [22] predicts 3D trajectories from each generated video, and performance is quantified by Average Multi-Object Tracking Accuracy (AMOTA) under the nuScenes protocol [9]. Higher AMOTA reflects more stable temporal dynamics and accurate data association for moving objects from the 3D scene.

D.4 **Occupancy Prediction** evaluates whether generated scenes enable accurate 3D reconstruction of spatial geometry and semantics. A frozen SparseOcc [96] predicts voxel-wise scene representations, and performance is measured using RayIoU, which compares semantic agreement along camera rays rather than volumetric overlap. Higher RayIoU indicates more accurate and depth-consistent occupancy estimation, showing that generated videos preserve 3D structural integrity crucial for downstream scene understanding.

3.5 Human Preference

This aspect evaluates alignment with **human judgment** by assessing how visually authentic, physically coherent, and behaviorally safe the generated videos appear to observers. Each dimension is rated on a '1' to '10' scale, where higher scores indicate stronger human perceptual fidelity.

H.1 **World Realism** measures the overall authenticity and naturalness of generated videos. We evaluate how closely textures, lighting, and motion resemble those in real-world driving footage. Three sub-dimensions are considered: (1) *Overall Realism*, capturing global scene coherence and natural appearance; (2) *Vehicle Realism*, assessing vehicle geometry, surface reflectance, and motion stability; and (3) *Pedestrian Realism*, focusing on body proportion and walking motion consistency. Higher scores indicate scenes that are visually indistinguishable from real recordings.

H.2 **Physical Plausibility** evaluates whether the scene obeys intuitive physical and causal principles. It focuses on the continuity of motion, correctness of occlusion order, object contact stability, and illumination consistency across time. Scenes exhibiting teleportation, interpenetration, or inconsistent reflections receive lower ratings, while those maintaining smooth transitions and physically coherent dynamics are scored higher.

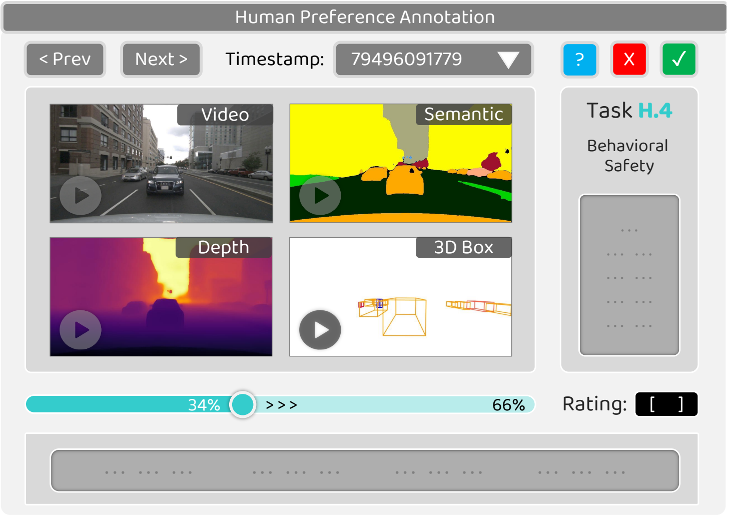
H.3 **3D & 4D Consistency** measures spatial-temporal coherence of geometry and appearance across frames. It assesses whether reconstructed 3D structures remain stable over time and whether objects preserve their relative positions, orientations, and trajectories. High consistency reflects that generated videos maintain realistic 3D layout and smooth temporal evolution, forming plausible 4D scenes aligned with real-world dynamics.

H.4 **Behavioral Safety** assesses whether traffic participants behave in predictable and risk-free ways consistent with common driving norms. It focuses on short-term interactions among vehicles, pedestrians, and environmental cues, such as adherence to traffic signals, collision avoidance, and stable lane following. Lower scores indicate abrupt or unsafe behaviors (*e.g.*, sudden collisions or unrealistic crossings), whereas higher scores correspond to smooth, lawful, and controlled motion indicative of safe and credible agent behavior.

4 Human Annotation & Evaluation Agent

4.1 Annotation Process

To establish reliable human supervision for our benchmark, we designed a structured multi-stage annotation pipeline. Ten annotators were divided into *two independent groups*, each responsible for scoring all videos under the four dimensions defined in Section 3.5. For every video and dimension, the two groups annotated separately; when their ratings diverged, the sample was re-evaluated to ensure consistency. Annotators were presented with **four synchronized views**: ^{1}*generated video*, ^{2}*semantic mask*, ^{3}*depth map*, and ^{4}*3D boxes*, through the interface shown in Figure 3. To promote consistency and domain understanding, all annotators received detailed documentation with examples illustrating each scoring level. On average, each annotation took approximately **two minutes and eight seconds**, amounting to over **930 hours**. Further implementation details, documentation, and examples are provided in the **Appendix**.


Timestamp: 79496091779

Rating: [ ]

Figure 3 Interface for **Human Preference** annotation process. We present four synchronized views: ^{1}generated video, ^{2}semantic mask, ^{3}depth map, and ^{4}3D bounding boxes, enabling comprehensive judgment of realism, physical plausibility, and consistency.

4.2 WorldLens-26K: A Diverse Preference Dataset

To bridge the gap between human judgment and automated evaluation, we curate a large-scale human-annotated dataset comprising **26,808 scoring records** of generated videos. Each entry includes a *discrete score* and a *concise textual rationale* written by annotators, capturing both quantitative assessment and qualitative explanation. The dataset covers complementary dimensions of perceptual quality (section 3.5). This balanced design ensures comprehensive coverage across spatial, temporal, and behavioral aspects of world-model realism. As shown in Figure 4, the word clouds of textual rationales align closely with their corresponding target dimensions, confirming the validity and interpretability of the collected labels. We envision **WorldLens-26K** as a foundational resource for training auto-evaluation agents and constructing human-aligned reward or advantage functions for reinforcement fine-tuning of generative world models.

Table 1 Benchmarking results of state-of-the-art driving world models for **Generation** and **Reconstruction** in WorldLens.

Model	**Venue**	**Aspect: Generation**								**Aspect: Reconstruction**			
		Subject Fid.↑ G.1	Subject Cohe.↑ G.2	Subject Cons.↑ G.3	Depth Disc.↓ G.4	Temp. Cons.↑ G.5	Sem. Cons.↑ G.6	Percept. Disc.↓ G.7	View Cons.↑ G.8	Photo. Error↓ R.1	Geo. Disc.↓ R.2	Novel Qual.↑ R.3	Novel Disc.↓ R.4
MagicDrive [27]	ICLR'23	28.49	75.95	65.22%	24.19	74.44%	80.63%	222.00	185.77	0.140	0.115	39.82%	427.30
DreamForge [69]	arXiv'24	31.99	75.12	76.40%	19.27	**79.82%**	84.99%	189.76	194.99	0.097	0.105	41.23%	347.70
DriveDreamer-2 [142]	AAAI'25	27.38	78.97	74.49%	17.73	79.51%	**85.91%**	127.07	302.83	0.093	**0.073**	36.10%	259.91
OpenDWM [89]	CVPR'25	**36.30**	**83.13**	**78.33%**	18.17	79.63%	84.08%	90.42	211.18	**0.065**	0.088	39.54%	287.73
DiST-4D [30]	ICCV'25	30.32	79.36	74.69%	**17.71**	77.76%	84.32%	**58.08**	**389.78**	0.066	0.080	**43.09%**	**192.39**
$\mathcal{X}$-Scene [128]	NeurIPS'25	27.17	77.22	74.37%	20.50	79.41%	83.80%	179.74	201.00	0.098	0.096	38.04%	365.71
Empirical Max	-	60.22	83.25	93.66%	14.27	93.24%	86.39%	-	570.75	0.056	-	45.69%	-

4.3 WorldLens-Agent: SFT from Human Feedback

Evaluating generated worlds hinges on human-centered criteria (*physical plausibility*) and subjective preferences (*perceived realism*) that quantitative metrics inherently miss, highlighting the necessity of a human-aligned evaluator. To this end, we introduce **WorldLens-Agent**, a vision-language critic agent trained on WorldLens-26K. Through LoRA-based supervised fine-tuning, we distill human perceptual and physical judgments into a Qwen3-VL-8B [2], enabling it to internalize criteria such as realism, plausibility, and behavioral safety. This provides consistent, human-aligned assessments, offering a scalable preference oracle for benchmarking future world models. Kindly refer to Figure 8 and the **Appendix** for automatic scoring and rationale generation examples on *out-of-distribution* videos.

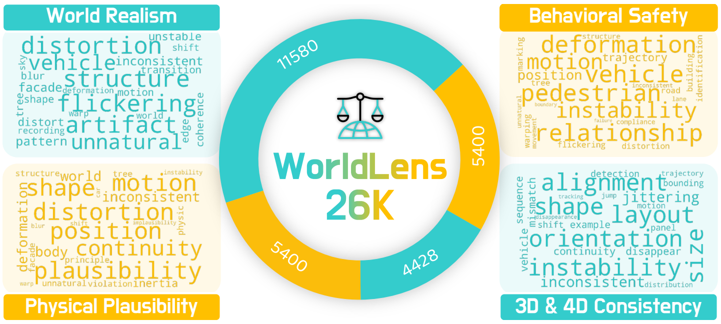


Figure 4 The statistics and word clouds of the **WorldLens-26K** dataset. Frequent keywords align closely with their target criteria (*e.g.*, "shape", "reflection", "motion", "safety"), confirming that annotators focus more on **dimension-specific perceptual attributes** and maintain consistent reasoning during the evaluation.

5 Experiments

We comprehensively evaluate representative driving world models across all five aspects defined in **WorldLens**, covering both quantitative and human-in-the-loop dimensions. Due to space constraints, detailed configurations, metrics, and implementation details are provided in the **Appendix**.

5.1 Per-Aspect Evaluations

Generation. As summarized in Table 1, all existing models remain notably below the 'Empirical Max', indicating substantial room for improving the visual and temporal realism of driving world models. Although DiST-4D [30] achieves the lowest *Perceptual Discrepancy*, it underperforms OpenDWM [89] in *Subject Fidelity* and *View Consistency*, demonstrating that perceptual metrics alone are insufficient for assessing physically coherent scene generation. OpenDWM provides the most balanced overall performance, largely due to large-scale multi-dataset training, while single-dataset models such as MagicDrive [27] and $\mathcal{X}$-Scene [128] exhibit limited generalization across all metrics. Notably, conditioned approaches like DiST-4D [30] and DriveDreamer-2 [142] partially overcome this limitation, improving *Depth* and *Cross-View Consistency* by 20–30% through the use of ground-truth frames. These results highlight that *the dataset diversity and conditioning strategies are more critical than perceptual fidelity for achieving reliable, temporally consistent world modeling.*

Reconstruction. We assess the spatiotemporal 3D coherence by reconstructing generated videos into 4D Gaussian Fields [17], where floaters and geometric instability directly reveal temporal inconsistency. As shown in Table 1, MagicDrive [27] exhibits the weakest reconstruction, with the highest *Photometric Error* and

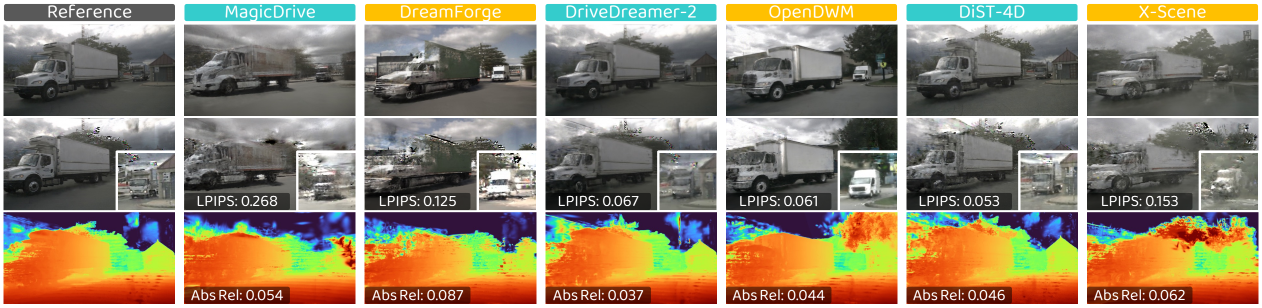


Figure 5 Qualitative results of **4D reconstruction** from generated videos. Rows (top to bottom) denote [1]generated frame, [2]rendered novel-view frame at a *Lateral Offset*, and [3]depth map. MagicDrive [27] and DreamForge [69] exhibit dense floaters and geometric distortions, while OpenDWM [89] and DiST-4D [30] maintain temporally more consistent geometry, aligning with the quantitative results in Table 1.

Geometric Discrepancy, both over **2**× worse than OpenDWM [89]. DreamForge [69] shows similar artifacts, indicating limited *3D Consistency*. In contrast, OpenDWM and DiST-4D [30] achieve markedly better reconstruction, reducing photometric and geometric errors by about 55% and producing more structurally coherent sequences. DiST-4D further attains the best *Novel-View Quality*, likely due to its RGB-D generation design that better preserves depth over time. As illustrated in Figure 5, MagicDrive and DreamForge produce dense floaters and distortions under lateral views, whereas OpenDWM and DiST-4D maintain clean, stable geometry. Overall, the results highlight that *the temporal stability and geometric consistency are essential for physically realistic and reconstructable world models.*

Action-Following. As shown in Table 2, we evaluate the functional viability of synthesized environments through closed-loop simulation, where a pretrained planner operates within the "world" of each model. Temporal coherence proves critical, as planning agents rely on multi-frame history and ego-state cues; models with weaker temporal stability achieve the lowest *Route Completion* rates. A notable finding is the large disparity between open-loop and closed-loop performance. Despite strong open-loop results on *Displacement Error* and *PDMS*, all methods collapse under closed-loop conditions, achieving only marginal *Route Completion* rates. Frequent failures (*e.g.*, collisions, off-road drift) suggest that current synthetic data remain inadequate substitutes for real-world data in high-level control. This highlights a key insight: *enhancing the physical and causal realism of generative world models is indispensable for effective closed-loop deployment.*

Table 2 Benchmarking results of state-of-the-art driving world models for **Action-Following** dimensions in WorldLens.

Model	Venue	Aspect: Action-Following			
		Displ. Error↓ A.1	Open-L Adh.↑ A.2	Route Compl.↑ A.3	Closed-L Adh.↑ A.4
MagicDrive [27]	ICLR'23	0.57	71.23%	6.89%	4.82%
Panacea [112]	CVPR'24	0.58	-	-	-
DreamForge [69]	arXiv'24	0.55	75.51%	10.23%	7.65%
DrivingSphere [121]	CVPR'25	0.54	76.02%	11.02%	8.29%
MagicDrive-V2 [28]	ICCV'25	**0.53**	**78.91%**	12.31%	9.50%
RLGF [120]	NeurIPS'25	**0.53**	78.45%	**13.51%**	**10.59%**
Empirical Max	-	0.51	-	-	-

Table 3 Summary of benchmarking results of state-of-the-art world models for **Downstream Task** dimensions in WorldLens.

Model	Venue	Aspect: Downstream Task			
		Map Seg.↑ D.1	3D Obj. Det.↑ D.2	3D Obj. Trk.↑ D.3	Occ. Pred.↑ D.4
MagicDrive [27]	ICLR'23	18.34%	22.41%	7.90%	23.14%
DreamForge [69]	arXiv'24	30.31%	26.71%	10.30%	23.71%
DriveDreamer-2 [142]	AAAI'25	33.62%	30.90%	13.30%	**26.82%**
OpenDWM [89]	CVPR'25	27.63%	21.96%	6.90%	24.82%
DiST-4D [30]	ICCV'25	**35.55%**	**33.22%**	**15.30%**	26.10%
$\mathcal{X}$-Scene [128]	NeurIPS'25	27.24%	29.89%	8.80%	23.68%
Empirical Max	-	40.64%	44.72%	36.30%	37.05%

Downstream Tasks. This aspect directly reflects the practical utility of world models beyond visual realism. As shown in Table 3 and Figure 1, DiST-4D [30] leads by a large margin across all tasks, (*i.e.*, map segmentation, 3D detection, and tracking), averaging 30–40% higher than the next best models. DriveDreamer-2 [142] ranks second, particularly excelling in occupancy prediction, highlighting the advantage of temporal conditioning for consistent

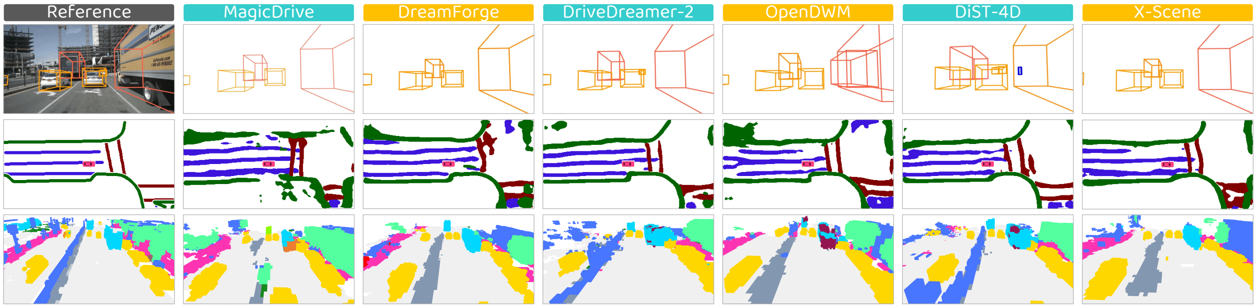


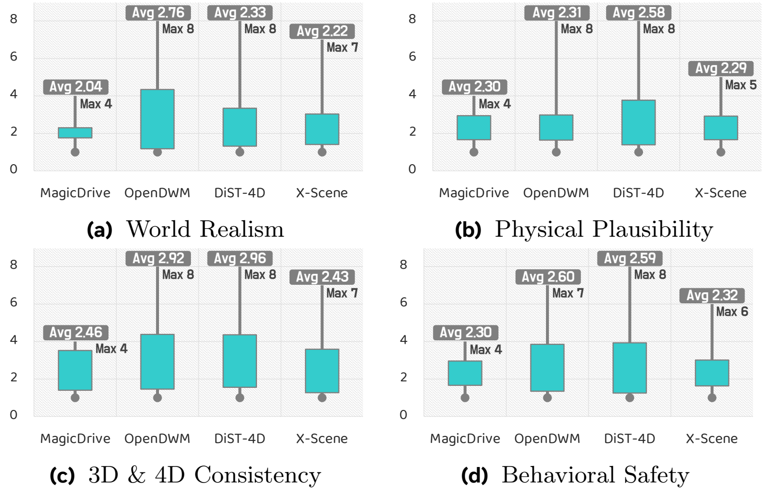
Figure 6 Qualitative results of **Downstream Tasks**. Rows (from top to bottom): [1]3D object detection, [2]map segmentation, and [3]semantic occupancy prediction tasks.

video generation. In contrast, MagicDrive [27] performs weakest across all tasks, confirming its limited spatiotemporal coherence. Interestingly, despite strong perceptual quality, OpenDWM [89] underperforms in detection (21.9%) and tracking (6.9%), suggesting that large-scale multi-domain training may hinder adaptation to specific dataset distributions. Additional qualitative assessments in Figure 6 further verify our observations. Overall, these results indicate that *the temporal conditioning and dataset alignment are critical for task-specific effectiveness for practical usages.*

5.2 Human Preference Alignments

Subjective Evaluations. Since not all aspects of world modeling can be captured by quantitative metrics, we conducted a human evaluation focusing on *World Realism*, *Physical Plausibility*, *3D & 4D Consistency*, and *Behavioral Safety*. As shown in Figure 7, overall scores remain modest (on average '2'∼'3' out of '10'), revealing that current world models are far from human-level realism. DiST-4D [30] achieves the most balanced scores across all dimensions, leading in physical plausibility ('2.58') and behavioral safety ('2.59'). OpenDWM [89] attains the highest realism ('2.76') but slightly lower physical consistency, while MagicDrive [27] ranks lowest overall, reflecting poor coherence. Interestingly, *World Realism* and *Consistency* scores correlate strongly, suggesting that human-perceived realism is tightly coupled with geometric and temporal stability. Overall, these results underscore the *necessity of human-in-the-loop evaluation to complement quantitative benchmarks and provide a holistic assessment of world-model quality.*

Human-Agent Alignments. To assess the generalizability of our automatic evaluator, **WorldLens-Agent**, we conduct a zero-shot test on videos generated by Gen3C [83], as shown in Figure 8. The agent's predicted scores exhibit strong alignment with human annotations across all evaluated dimensions, confirming its ability to capture nuanced subjective preferences. Beyond numerical agreement, the textual rationales generated by the agent closely mirror those written by human annotators, demonstrating both *score-level consistency* and *interpretive coherence*. These results highlight the effectiveness of leveraging human-annotated perception data to train scalable, explainable, and reproducible evaluative agents for future world-model benchmarking.



(a) World Realism

(b) Physical Plausibility

(c) 3D & 4D Consistency

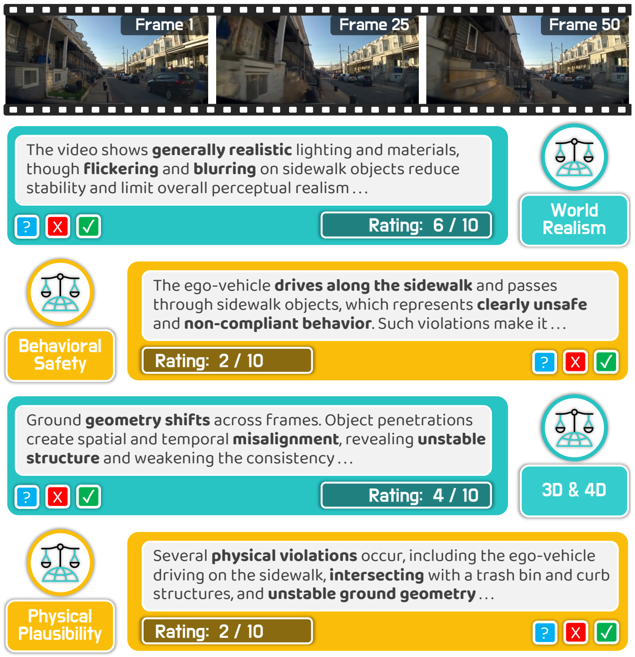
(d) Behavioral Safety

Figure 7 Summary of alignments to **Human Preference**, where the max, median, and average scores of each model are compared. For more detailed analyses, kindly refer to the **Appendix**.

5.3 Insights & Discussions

Comprehensive Evaluations are Crucial. No single world model excels in all aspects (Figure 1): DiST-4D performs best in geometry and novel-view metrics, OpenDWM leads in photometric fidelity, and DriveDreamer-2 achieves the highest depth accuracy. This divergence shows that visual realism, geometric consistency, and downstream usability are *complementary rather than interchangeable, highlighting the necessity of multi-dimensional benchmarking.*

Perceptual Quality Does Not Imply Usability. Models with strong perceptual scores (*e.g.*, OpenDWM) may underperform on downstream tasks. Despite its visual fidelity, OpenDWM scores *30% lower* than DiST-4D in 3D detection, indicating that large-scale, multi-domain training can hinder adaptation to task-specific distributions. Hence, *aligning generated data with the target domain is more crucial than perceptual realism for effective downstream use.*



The video shows **generally realistic** lighting and materials, though **flickering** and **blurring** on sidewalk objects reduce stability and limit overall perceptual realism ...

Rating: 6 / 10

The ego-vehicle **drives along the sidewalk** and passes through sidewalk objects, which represents **clearly unsafe** and **non-compliant behavior**. Such violations make it ...

Rating: 2 / 10

Ground **geometry shifts** across frames. Object penetrations create spatial and temporal **misalignment**, revealing **unstable structure** and weakening the consistency ...

Rating: 4 / 10

Several **physical violations** occur, including the ego-vehicle driving on the sidewalk, **intersecting** with a trash bin and curb structures, and **unstable ground geometry** ...

Rating: 2 / 10

Figure 8 Zero-shot evaluations by **WorldLens-Agent** on unseen videos (from Gen3C [83]), exhibiting strong alignments with human reasoning. See the **Appendix** for more examples.

Geometry Awareness Enables Physical Coherence. The superior reconstruction and novel-view performance of DiST-4D stem from its RGB-D generation and decoupled spatiotemporal diffusion, which jointly model temporal forecasting and spatial synthesis. This shows that *geometry-aware supervision significantly improves the physical realism and reconstructability of generated scenes.*

Joint Optimization of Appearance and Geometry. The discrepancy between photometric (LPIPS/PSNR) and geometric metrics (Abs Rel) reveals that current models often treat texture and structure as independent objectives. Geometry-aware supervision stabilizes depth but blurs details, while appearance-driven training sharpens textures yet breaks spatial consistency. A unified formulation that *jointly optimizes appearance and geometry through spatiotemporal regularization yields consistent reconstruction.*

Guidelines for Future World Model Design. Key principles emerge for developing physically grounded world models: *1) Prioritize geometry as a core objective:* explicit depth prediction or supervision enhances both reconstruction fidelity and downstream perception; *2) Stabilize foreground dynamics:* consistent geometry is essential for reliable motion disentanglement; *3) Ensure autoregressive resilience:* enforcing cross-view and temporal consistency mitigates drift and structural artifacts, while training with self-forcing [18, 40] or streaming diffusion [51] enhances robustness against compounding errors in closed-loop generation, which is crucial for long-horizon stability. Overall, robust world models stem from the *joint optimization of appearance, geometry, and task adaptability*, advancing from *visual realism* toward *physical reliability.*

6 Conclusion

We presented **WorldLens**, a full-spectrum benchmark that evaluates generative world models across perception, geometry, function, and human alignment perspectives. Through five complementary evaluation aspects and over twenty standardized metrics, it offers a unified protocol for measuring both physical and perceptual realism. Together with **WorldLens-26K** and **WorldLens-Agent**, our framework establishes a scalable, interpretable foundation for benchmarking future world models – guiding progress toward systems that not only *look* real but also *behave* realistically.

Appendix

7 Aspect 1: Generation

In this section, we detail the metrics used to evaluate the **quality of generation** of driving world models. This aspect assesses the overall realism, coherence, and physical plausibility of generated driving videos, capturing how well a model reconstructs the spatiotemporal structure of real-world scenes.

7.1 Subject Fidelity

7.1.1 Definition

Subject Fidelity quantifies the perceptual realism of object instances, such as vehicles and pedestrians, that appear in generated driving videos. It focuses on assessing whether each synthesized object visually resembles its real-world counterpart in both appearance and semantic attributes. By isolating individual instances, this metric emphasizes fine-grained visual fidelity that global perceptual measures may overlook, providing an object-centric view of generation quality.

7.1.2 Formulation

For a generated video $y_j = \{y_j^{(t)}\}_{t=1}^{T}$ with bounding boxes $\{b_{j,k}^{(t)}\}_{k=1}^{K_j^{(t)}}$, we crop object patches $o_{j,k}^{(t)} = \mathrm{Crop}(y_j^{(t)}, b_{j,k}^{(t)})$. Let $\mathcal{C}$ denote the evaluated object categories (*e.g.*, vehicle, pedestrian), and $\psi_{\mathrm{CLS}}^{(c)}(\cdot)$ be a pretrained binary classifier for class $c \in \mathcal{C}$ outputting confidence $p_{j,k}^{(t,c)} \in [0, 1]$ that patch $o_{j,k}^{(t)}$ looks real for that class. Aggregating across all objects, frames, videos, and classes yields the overall Subject Fidelity score:

$$\boxed{\mathcal{S}_{\mathrm{SF}}(\mathcal{Y}) = \frac{1}{N_g|\mathcal{C}|} \sum_{j=1}^{N_g} \sum_{c\in\mathcal{C}} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{K_{j,c}^{(t)}} \sum_{k=1}^{K_{j,c}^{(t)}} p_{j,k}^{(t,c)}} \tag{1}$$

A higher $\mathcal{S}_{\mathrm{SF}}$ score indicates that generated objects are both visually convincing and semantically consistent with their intended categories. Models achieving high fidelity tend to produce realistic textures, shapes, and colors, even under varying viewpoints and lighting conditions. This metric thus complements global measures like FVD or LPIPS by focusing on localized realism at the instance level, offering insights into whether the generated world contains physically believable and semantically meaningful entities.

7.1.3 Implementation Details

We use class-specific confidence scores for evaluation. Pedestrian crops are classified using a pedestrian classifier pretrained on several commonly used pedestrian-datasets [19, 76, 110, 144], while vehicle crops are classified with a ViT-B/16 model (‘`google/vit-base-patch16-224`‘) [113] pretrained on ImageNet-21k (14 million images, 21,843 classes) [21]. Category grouping is determined by regex-based label matching against the model’s `id2label`. Images are resized to 256 × 128 and normalized before inference. For each tracklet, we average the classification confidence of all selected frames, and report the mean confidence as the final score.

7.1.4 Examples

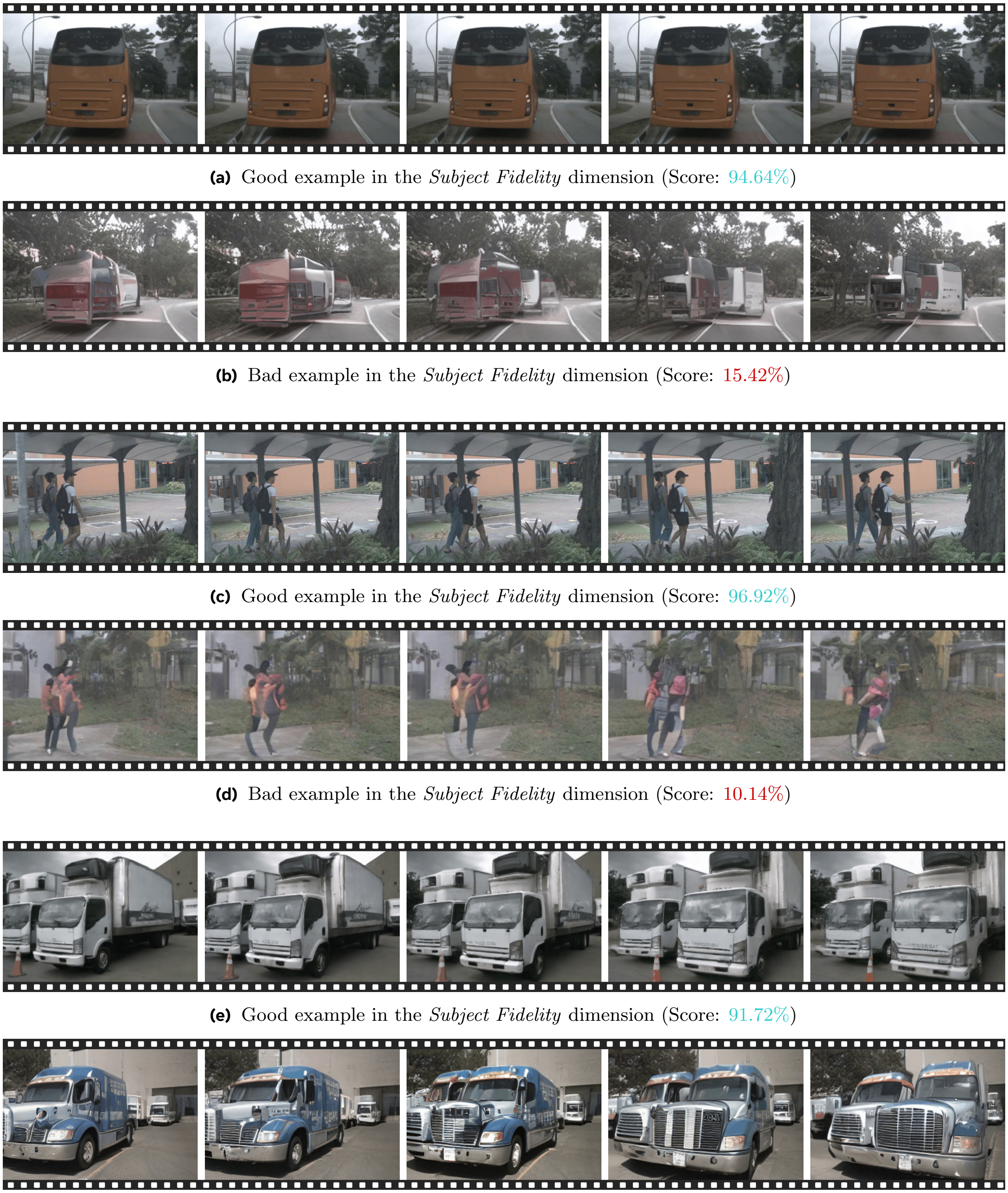
Figure 9 provides typical examples of videos with good and bad quality in terms of *Subject Fidelity*.

7.1.5 Evaluation & Analysis

Table 4 provides the complete results of models in terms of *Subject Fidelity*.

Table 4 Complete results of state-of-the-art driving world models in terms of *Subject Fidelity* in WorldLens.

$\mathcal{S}_{\mathrm{SF}}(\cdot)$	**MagicDrive**	**DreamForge**	**DriveDreamer-2**	**OpenDWM**	**DiST-4D**	**$\mathcal{X}$-Scene**	**Empirical**
	[ICLR’24]	[arXiv’24]	[AAAI’25]	[CVPR’25]	[ICCV’25]	[NeurIPS’25]	**Max**
Vehicle (↑)	26.16%	26.97%	24.23%	34.04%	28.02%	24.03%	56.10%
Pedestrian (↑)	49.45%	77.13%	55.69%	56.65%	50.98%	55.42%	97.27%
Total (↑)	28.49%	31.99%	27.38%	36.30%	30.32%	27.17%	60.22%



(a) Good example in the *Subject Fidelity* dimension (Score: 94.64%)

(b) Bad example in the *Subject Fidelity* dimension (Score: 15.42%)

(c) Good example in the *Subject Fidelity* dimension (Score: 96.92%)

(d) Bad example in the *Subject Fidelity* dimension (Score: 10.14%)

(e) Good example in the *Subject Fidelity* dimension (Score: 91.72%)

(f) Bad example in the *Subject Fidelity* dimension (Score: 41.75%)

Figure 9 Examples of "good" and "bad" generation qualities in terms of *Subject Fidelity* in WorldLens.

7.2 Subject Coherence

7.2.1 Definition

Subject Coherence evaluates the temporal stability of an object's visual identity across consecutive frames within a generated sequence. It captures whether the same entity – such as a specific car or pedestrian – maintains consistent appearance attributes, including color, texture, and shape, over time. This metric assesses not only visual continuity but also the preservation of object identity, which is crucial for generating physically plausible and temporally coherent scenes for autonomous driving applications.

7.2.2 Formulation

For each generated video $y_j = \{y_j^{(t)}\}_{t=1}^{T}$, the conditioning provides bounding boxes $\{b_{j,k}^{(t)}\}$ and associated track IDs $r_{j,k}$. Object patches are cropped as $o_{j,r}^{(t)} = \mathrm{Crop}(y_j^{(t)}, b_{j,k}^{(t)})$ for object track $r = r_{j,k}$. A frozen ReID encoder $\phi_{\mathrm{ReID}}(\cdot)$ extracts ℓ_2-normalized embeddings:

$$\mathbf{g}_{j,r}^{(t)} = \phi_{\mathrm{ReID}}\left(o_{j,r}^{(t)}\right), \qquad \|\mathbf{g}_{j,r}^{(t)}\|_2 = 1.$$

The dataset-level Subject Coherence is computed as the mean cosine similarity between consecutive embeddings of the same tracked object, aggregated over all tracks, frames, and videos:

$$\boxed{\mathcal{S}_{\mathrm{SC}}(\mathcal{Y}) = \tfrac{1}{N_g}\sum\nolimits_{j=1}^{N_g} \tfrac{1}{R_j}\sum\nolimits_{r=1}^{R_j} \tfrac{1}{T_r-1}\sum\nolimits_{t=1}^{T_r-1} \mathbf{g}_{j,r}^{(t)\top}\mathbf{g}_{j,r}^{(t+1)}} \tag{2}$$

where R_j is the number of track IDs in video y_j and T_r the number of frames where object r appears. A high $\mathcal{S}_{\mathrm{SC}}$ score reflects consistent and temporally stable object generation, indicating that the model preserves identity-related features despite changes in position, viewpoint, or lighting. In contrast, a low score often signals flickering textures, shape distortions, or identity switches between frames.

This metric thus serves as a sensitive indicator of temporal realism, distinguishing models that produce temporally coherent scenes from those limited to frame-wise synthesis.

7.2.3 Implementation Details

We compute *Subject Coherence* using embeddings extracted from the Cross-Video ReID model of Zuo et al. [148]. Frames are filtered using confidence thresholds of 0.25 for vehicles and 0.50 for pedestrians before similarity computation. The final score is a combination of both sub-metrics.

7.2.4 Examples

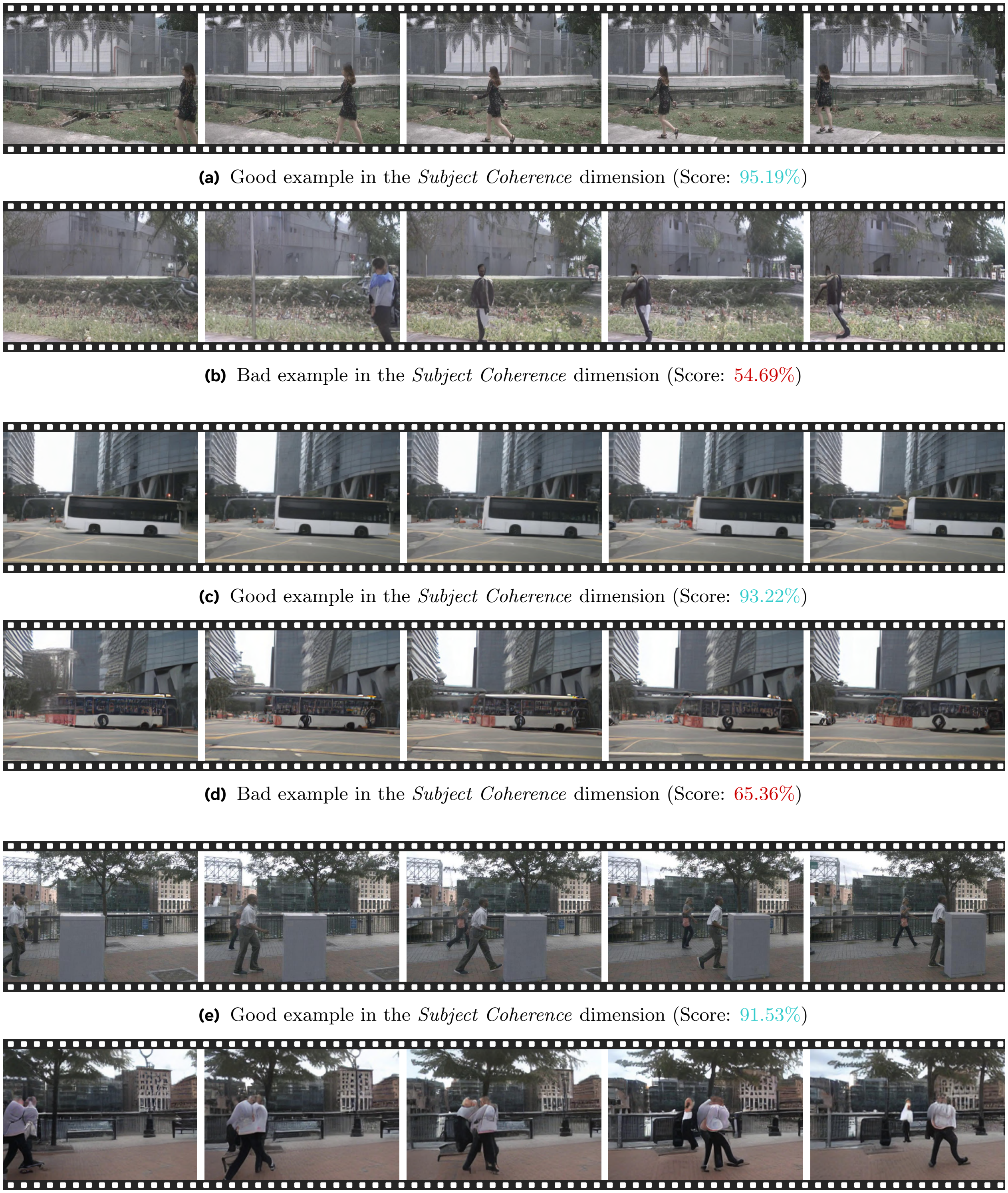
Figure 10 provides typical examples of videos with good and bad quality in terms of *Subject Coherence*.

7.2.5 Evaluation & Analysis

Table 5 provides the complete results of models in terms of *Subject Coherence*.

Table 5 Complete results of state-of-the-art driving world models in terms of *Subject Coherence* in WorldLens.

$\mathcal{S}_{\mathrm{SC}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical** **Max**
Vehicle (↑)	72.12%	72.00%	77.45%	82.03%	78.51%	74.02%	82.86%
Pedestrian (↑)	79.78%	78.23%	80.48%	84.22%	80.20%	80.42%	83.25%
Total (↑)	75.95%	75.12%	78.97%	83.13%	79.36%	77.22%	83.25%



(a) Good example in the *Subject Coherence* dimension (Score: 95.19%)

(b) Bad example in the *Subject Coherence* dimension (Score: 54.69%)

(c) Good example in the *Subject Coherence* dimension (Score: 93.22%)

(d) Bad example in the *Subject Coherence* dimension (Score: 65.36%)

(e) Good example in the *Subject Coherence* dimension (Score: 91.53%)

(f) Bad example in the *Subject Coherence* dimension (Score: 66.79%)

Figure 10 Examples of “good” and “bad” generation qualities in terms of *Subject Coherence* in WorldLens.

7.3 Subject Consistency

7.3.1 Definition

Subject Consistency measures stability of object-level semantics and structural details. It focuses on fine-grained appearance and geometric regularity through DINO features [12], evaluating whether dynamic subjects maintain consistent texture, shape, and structure over time. High scores indicate that the model preserves the semantic identity and visual integrity of objects throughout motion, avoiding flickering or deformation.

7.3.2 Formulation

For each generated video $y_j = \{y_j^{(t)}\}_{t=1}^{T}$ and its paired ground-truth $x_j = \{x_j^{(t)}\}_{t=1}^{T}$, we extract ℓ_2-normalized DINO embeddings: $\mathbf{g}_j^{(t)} = \phi_{\text{DINO}}(y_j^{(t)})$, $\mathbf{f}_j^{(t)} = \phi_{\text{DINO}}(x_j^{(t)})$, and $\|\mathbf{g}_j^{(t)}\|_2 = \|\mathbf{f}_j^{(t)}\|_2 = 1$, where $\phi_{\text{DINO}}(\cdot)$ denotes the frozen DINO feature extractor. To quantify temporal stability, we compute three complementary terms:

- **Adjacent-Frame Smoothness:**

$$\text{ACM}(y_j) = \tfrac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{g}_j^{(t)\top} \mathbf{g}_j^{(t+1)} \ ,$$

  which measures the average cosine similarity between consecutive frame embeddings.

- **Temporal Jitter Index (TJI):**

$$\text{TJI}(y_j) = \tfrac{1}{T-2} \sum_{t=2}^{T-1} \frac{\|\mathbf{g}_j^{(t+1)} - 2\mathbf{g}_j^{(t)} + \mathbf{g}_j^{(t-1)}\|_2}{\frac{1}{2}\big(\|\mathbf{g}_j^{(t+1)} - \mathbf{g}_j^{(t)}\|_2 + \|\mathbf{g}_j^{(t)} - \mathbf{g}_j^{(t-1)}\|_2\big) + \varepsilon} \ ,$$

  which measures normalized second-order fluctuations (lower is smoother).

- **Motion-Rate Similarity (MRS):**

$$\text{MRS}(y_j, x_j) = \exp\Big(-\beta \, \tfrac{1}{T-1} \sum_{t=1}^{T-1} \big| \log \tfrac{\|\mathbf{g}_j^{(t+1)} - \mathbf{g}_j^{(t)}\|_2 + \varepsilon}{\|\mathbf{f}_j^{(t+1)} - \mathbf{f}_j^{(t)}\|_2 + \varepsilon} \big| \Big) \ ,$$

  which aligns the per-frame feature motion magnitude with that of the ground-truth sequence.

The overall Subject Consistency score integrates these terms:

$$\boxed{\mathcal{S}_{\text{SC}}(\mathcal{Y}) = \tfrac{1}{N_g} \sum_{j=1}^{N_g} \frac{\text{ACM}(y_j)}{1 + \text{TJI}(y_j)} \cdot \text{MRS}(y_j, x_j)^{1/2}} \tag{3}$$

7.3.3 Implementation Details

We extract frame-wise features using DINO ViT-B/16 [12]. These embeddings are used to compute adjacent-frame similarity, temporal jitter, and motion alignment against the corresponding ground-truth videos.

7.3.4 Examples

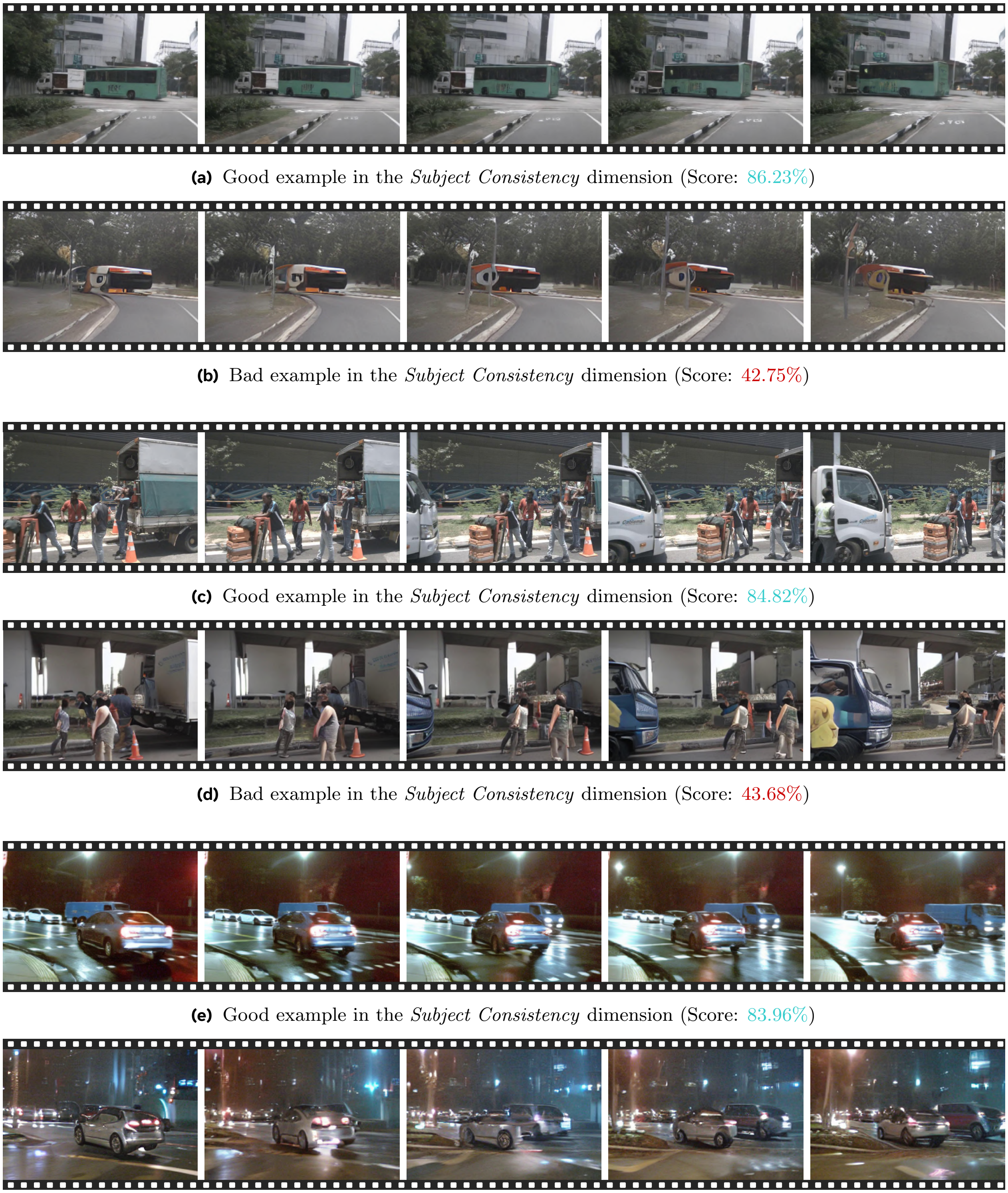
Figure 11 provides typical examples of videos with good and bad quality in terms of *Subject Consistency*.

7.3.5 Evaluation & Analysis

Table 6 provides the complete results of models in terms of *Subject Consistency*.

Table 6 Complete results of state-of-the-art driving world models in terms of *Subject Consistency* in WorldLens.

$\mathcal{S}_{\text{SC}}(\cdot)$	**MagicDrive**	**DreamForge**	**DriveDreamer-2**	**OpenDWM**	**DiST-4D**	**$\mathcal{X}$-Scene**	**Empirical**
	[ICLR'24]	[arXiv'24]	[AAAI'25]	[CVPR'25]	[ICCV'25]	[NeurIPS'25]	**Max**
ACM (↑)	89.32%	91.72%	90.09%	92.21%	91.15%	90.72%	93.66%
TJI (↑)	44.12%	43.32%	45.37%	44.95%	45.79%	43.41%	45.94%
Total (↑)	65.22%	76.40%	74.49%	78.33%	74.69%	74.37%	93.66%



(a) Good example in the *Subject Consistency* dimension (Score: 86.23%)

(b) Bad example in the *Subject Consistency* dimension (Score: 42.75%)

(c) Good example in the *Subject Consistency* dimension (Score: 84.82%)

(d) Bad example in the *Subject Consistency* dimension (Score: 43.68%)

(e) Good example in the *Subject Consistency* dimension (Score: 83.96%)

(f) Bad example in the *Subject Consistency* dimension (Score: 42.53%)

Figure 11 Examples of “good” and “bad” generation qualities in terms of *Subject Consistency* in WorldLens.

7.4 Depth Discrepancy

7.4.1 Definition

Depth Discrepancy quantifies the temporal stability of depth representations inferred from generated video sequences. In natural driving scenes, the apparent depth of foreground and background objects evolves smoothly with camera motion, whereas inconsistent generation often introduces discontinuous jumps in predicted depth. This metric captures such instability by measuring temporal variation in depth embeddings extracted from consecutive frames, providing a geometric complement to perceptual fidelity metrics.

7.4.2 Formulation

For a generated video $y_j = \{y_j^{(t)}\}_{t=1}^{T}$, we estimate per-frame depth maps using a monocular depth estimator $\psi_{\text{Depth}}(\cdot)$:

$$d_j^{(t)} = \psi_{\text{Depth}}\left(y_j^{(t)}\right), \qquad d_j^{(t)} \in \mathbb{R}^{H\times W}.$$

Each depth map is RGB-encoded by a fixed colormap $\mathcal{C}$ and processed by a pretrained visual encoder $\phi_{\text{DINO}}(\cdot)$ to obtain global embeddings:

$$f_j^{(t)} = \phi_{\text{DINO}}\left(\mathcal{C}(d_j^{(t)})\right), \qquad f_j^{(t)} \in \mathbb{R}^{D}.$$

Temporal variation in depth representation is then measured by the mean L2 distance between consecutive embeddings:

$$\text{DD}(y_j) = \tfrac{1}{T-1}\sum\nolimits_{t=1}^{T-1} \|f_j^{(t)} - f_j^{(t+1)}\|_2 \ .$$

Finally, the dataset-level Depth Discrepancy can be calculated as follows:

$$\boxed{\mathcal{S}_{\text{DD}}(\mathcal{Y}) = \tfrac{1}{N_g}\sum\nolimits_{j=1}^{N_g} \text{DD}(y_j)} \tag{4}$$

Lower $\mathcal{S}_{\text{Depth}}$ indicates smoother, more physically consistent depth evolution across time, reflecting stronger temporal geometric stability in the generated videos.

7.4.3 Implementation Details

Depth maps for both generated and ground-truth videos are obtained using Video DepthAnything [15]. The predicted depths are directly used to compute the per-frame depth discrepancy.

7.4.4 Examples

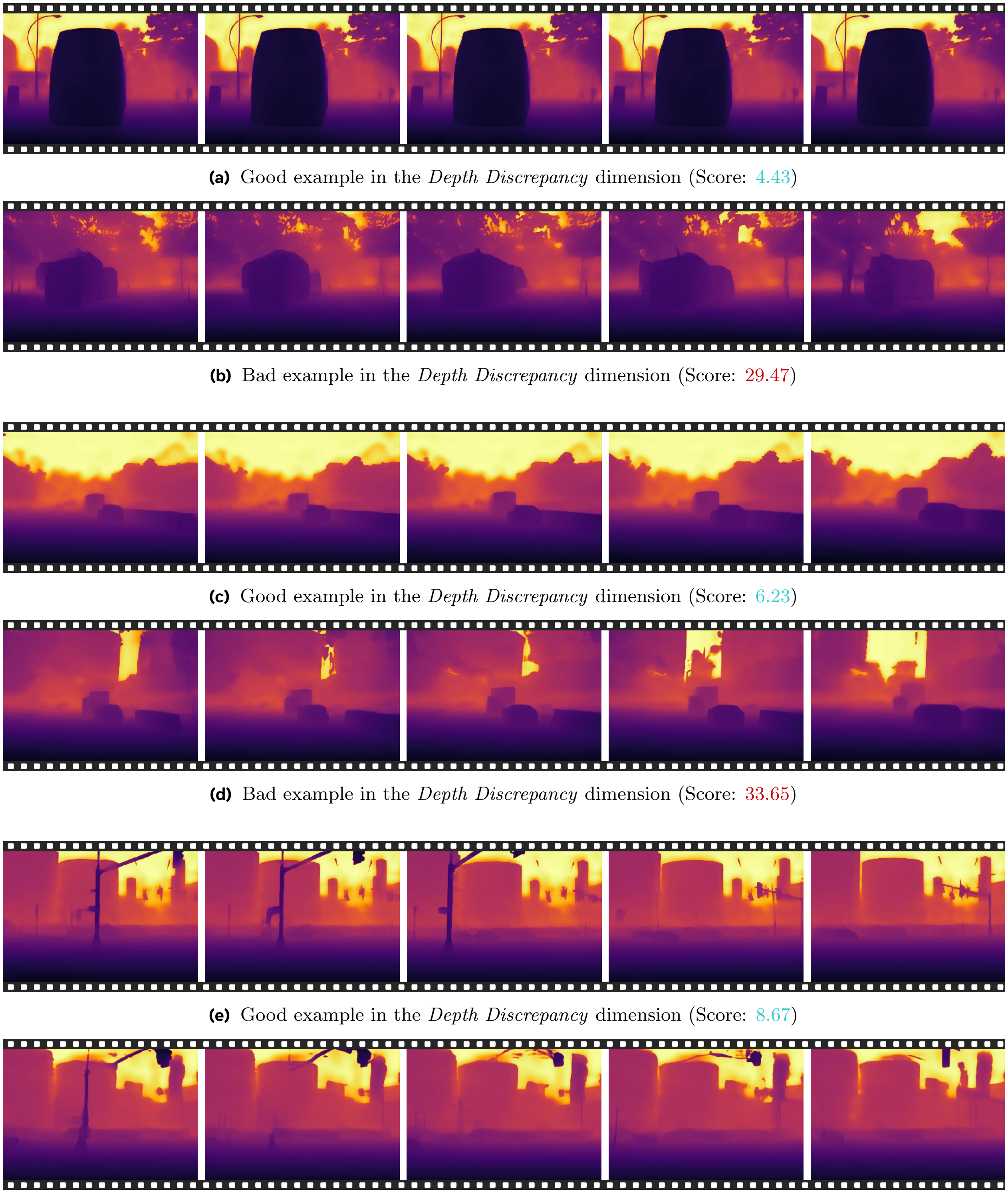
Figure 12 provides typical examples of videos with good and bad quality in terms of *Depth Discrepancy.*

7.4.5 Evaluation & Analysis

Table 7 provides the complete results of models in terms of *Depth Discrepancy.*

Table 7 Complete comparisons of state-of-the-art driving world models in terms of *Depth Discrepancy* in WorldLens.

$\mathcal{S}_{\text{DD}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
Total (↓)	24.19	19.27	17.73	18.17	17.71	20.50	14.27



(a) Good example in the *Depth Discrepancy* dimension (Score: 4.43)

(b) Bad example in the *Depth Discrepancy* dimension (Score: 29.47)

(c) Good example in the *Depth Discrepancy* dimension (Score: 6.23)

(d) Bad example in the *Depth Discrepancy* dimension (Score: 33.65)

(e) Good example in the *Depth Discrepancy* dimension (Score: 8.67)

(f) Bad example in the *Depth Discrepancy* dimension (Score: 19.34)

Figure 12 Examples of “good” and “bad” generation qualities in terms of *Depth Discrepancy* in WorldLens.

7.5 Temporal Consistency

7.5.1 Definition

Temporal Consistency quantifies the frame-to-frame stability of generated videos in a learned appearance space. Using a frozen CLIP encoder, this metric captures whether visual representations evolve smoothly over time without abrupt changes or flickering. It measures: (1) adjacent-frame smoothness, (2) suppression of high-frequency temporal jitter, and (3) alignment of motion magnitudes with real sequences. Together, these components evaluate whether generated videos exhibit physically coherent and temporally realistic dynamics.

7.5.2 Formulation

For each generated video $y_j = \{y_j^{(t)}\}_{t=1}^{T}$ and its paired ground-truth $x_j = \{x_j^{(t)}\}_{t=1}^{T}$, we extract ℓ_2-normalized CLIP embeddings as follows:

$$\mathbf{g}_j^{(t)} = \phi_{\text{CLIP}}\left(y_j^{(t)}\right), \qquad \mathbf{f}_j^{(t)} = \phi_{\text{CLIP}}\left(x_j^{(t)}\right), \qquad \|\mathbf{g}_j^{(t)}\|_2 = \|\mathbf{f}_j^{(t)}\|_2 = 1.$$

Follow the temporal statistics calculations in *Subject Consistency* (Eq. 7.3), the adjacent-frame smoothness, jitter suppression, and motion-rate alignment are applied in this CLIP space. Combining these components, the per-video score is defined as:

$$\text{TC}(y_j) = \tfrac{\text{ACM}(y_j)}{1+\text{TJI}(y_j)}\,\text{MRS}(y_j, x_j)^{1/2}.$$

The dataset-level metric averages per-video scores:

$$\boxed{\mathcal{S}_{\text{TC}}(\mathcal{Y}) = \tfrac{1}{N_g}\sum_{j=1}^{N_g}\text{TC}(y_j)} \tag{5}$$

with $\varepsilon = 10^{-8}$ and $\beta = 0.5$. By construction, $\text{ACM} \in [0, 1]$ and $\text{TJI} \geq 0$.

A high $\mathcal{S}_{\text{TC}}$ score indicates that appearance features change gradually across frames, producing smooth motion and physically coherent dynamics. Low scores correspond to flickering, abrupt illumination shifts, or motion discontinuities. This metric captures the degree to which generated sequences maintain continuity in both content and motion, serving as a robust proxy for temporal realism in driving videos.

7.5.3 Implementation Details

Temporal Consistency is evaluated using frame-wise features from CLIP ViT-B/32 [79] with an input resolution of 224 × 224. The normalized embeddings are used to derive adjacent-frame similarity, a temporal jitter index, and motion alignment between generated and ground-truth videos.

7.5.4 Examples

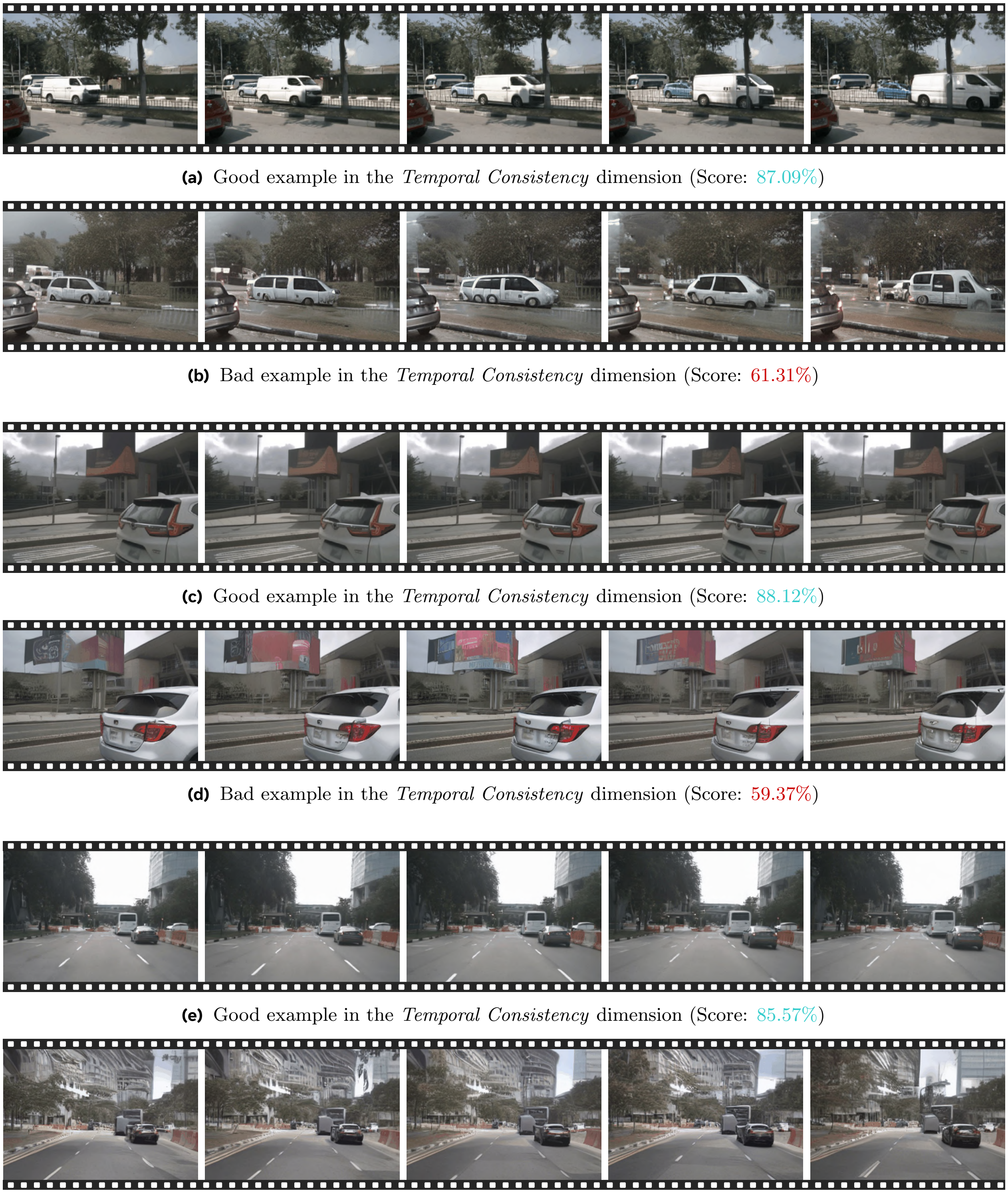
Figure 13 provides typical examples of videos with good and bad quality in terms of *Temporal Consistency*.

7.5.5 Evaluation & Analysis

Table 8 provides the complete results of models in terms of *Temporal Consistency*.

Table 8 Complete comparisons of state-of-the-art driving world models in terms of *Temporal Consistency* in WorldLens.

$\mathcal{S}_{\text{TC}}(\cdot)$	**MagicDrive**	**DreamForge**	**DriveDreamer-2**	**OpenDWM**	**DiST-4D**	**$\mathcal{X}$-Scene**	**Empirical Max**
	[ICLR'24]	[arXiv'24]	[AAAI'25]	[CVPR'25]	[ICCV'25]	[NeurIPS'25]	
ACM (↑)	91.43%	92.69%	93.65%	93.55%	92.27%	92.26%	93.24%
TJI (↑)	43.31%	42.69%	44.19%	43.83%	44.22%	42.91%	45.87%
Total (↑)	74.44%	79.82%	79.51%	79.63%	77.76%	79.41%	93.24%



(a) Good example in the *Temporal Consistency* dimension (Score: 87.09%)

(b) Bad example in the *Temporal Consistency* dimension (Score: 61.31%)

(c) Good example in the *Temporal Consistency* dimension (Score: 88.12%)

(d) Bad example in the *Temporal Consistency* dimension (Score: 59.37%)

(e) Good example in the *Temporal Consistency* dimension (Score: 85.57%)

(f) Bad example in the *Temporal Consistency* dimension (Score: 54.45%)

Figure 13 Examples of "good" and "bad" generation qualities in terms of *Temporal Consistency* in WorldLens.

7.6 Semantic Consistency

7.6.1 Definition

Semantic Consistency assesses the temporal stability of scene semantics in generated videos, ensuring that the underlying segmentation layout evolves smoothly over time. Using a frozen semantic segmentation model $\psi_{\text{SEG}}(\cdot)$, this metric quantifies how consistently pixel-wise labels, region structures, and global class distributions are preserved between consecutive frames.

7.6.2 Formulation

For each generated video $y_j = \{y_j^{(t)}\}_{t=1}^{T}$, we obtain frame-wise segmentation masks: $M_j^{(t)} = \psi_{\text{SEG}}(y_j^{(t)}) \in \{0, \dots, C{-}1\}^{H \times W}$. Temporal semantic stability is quantified by three complementary components:

Label Flip Rate (LFR) measures how rarely *interior* pixels (after class-wise morphological erosion) change their semantic label between consecutive frames. For class c, let $\Omega_c^{(t)}$ be the eroded interior region. The flip ratio is the fraction of pixels in $\Omega_c^{(t)}$ whose labels differ in $M_j^{(t+1)}$. The per-video LFR score averages these values across classes and time, then normalizes: $\mathcal{S}_{\text{LFR}}(y_j) = 1 - \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{\sum_c \sum_{\mathbf{p} \in \Omega_c^{(t)}} \mathbf{1}[M_j^{(t+1)}(\mathbf{p}) \neq c]}{\sum_c |\Omega_c^{(t)}|}$.

Segment Association Consistency (SAC) measures how consistently connected semantic regions persist over time. For each class c, connected components in $M_j^{(t)}$ and $M_j^{(t+1)}$ are matched by Hungarian assignment over IoU. The score is the pixel-weighted mean IoU of the matched region pairs: $\mathcal{S}_{\text{SAC}}(y_j) = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{\sum_c \sum_{(R,R') \in \pi_c^{(t)}} |R| \cdot \text{IoU}(R,R')}{\sum_c \sum_{R \in \mathcal{R}_c^{(t)}} |R|}$, where $\pi_c^{(t)}$ is the optimal region matching.

Class Distribution Stability (CDS) compares frame-level class histograms. Let $p^{(t)}$ be the normalized histogram of frame t. Global distribution shift is quantified by the Jensen–Shannon divergence: $\mathcal{S}_{\text{CDS}}(y_j) = 1 - \frac{1}{T-1} \sum_{t=1}^{T-1} \text{JSD}\big(p^{(t)} \,\|\, p^{(t+1)}\big)$.

Each component is normalized to $[0, 1]$. The final Semantic Consistency score is a weighted combination:

$$\boxed{\mathcal{S}_{\text{SemC}}(\mathcal{Y}) = \frac{1}{N_g} \sum_{j=1}^{N_g} \big[w_1 \mathcal{S}_{\text{LFR}}(y_j) + w_2 \mathcal{S}_{\text{SAC}}(y_j) + w_3 \mathcal{S}_{\text{CDS}}(y_j)\big]} \quad (6)$$

with $(w_1, w_2, w_3) = (0.5, 0.4, 0.1)$. A high $\mathcal{S}_{\text{SemC}}$ score signifies that drivable areas, lane boundaries, and object classes remain stable under temporal changes.

7.6.3 Implementation Details

We obtain frame-wise semantic maps using the panoptic segmentation model from OpenSeeD [138]. The predicted segments are then converted to label masks via a fixed color palette and used to compute the temporal semantic consistency score.

7.6.4 Examples

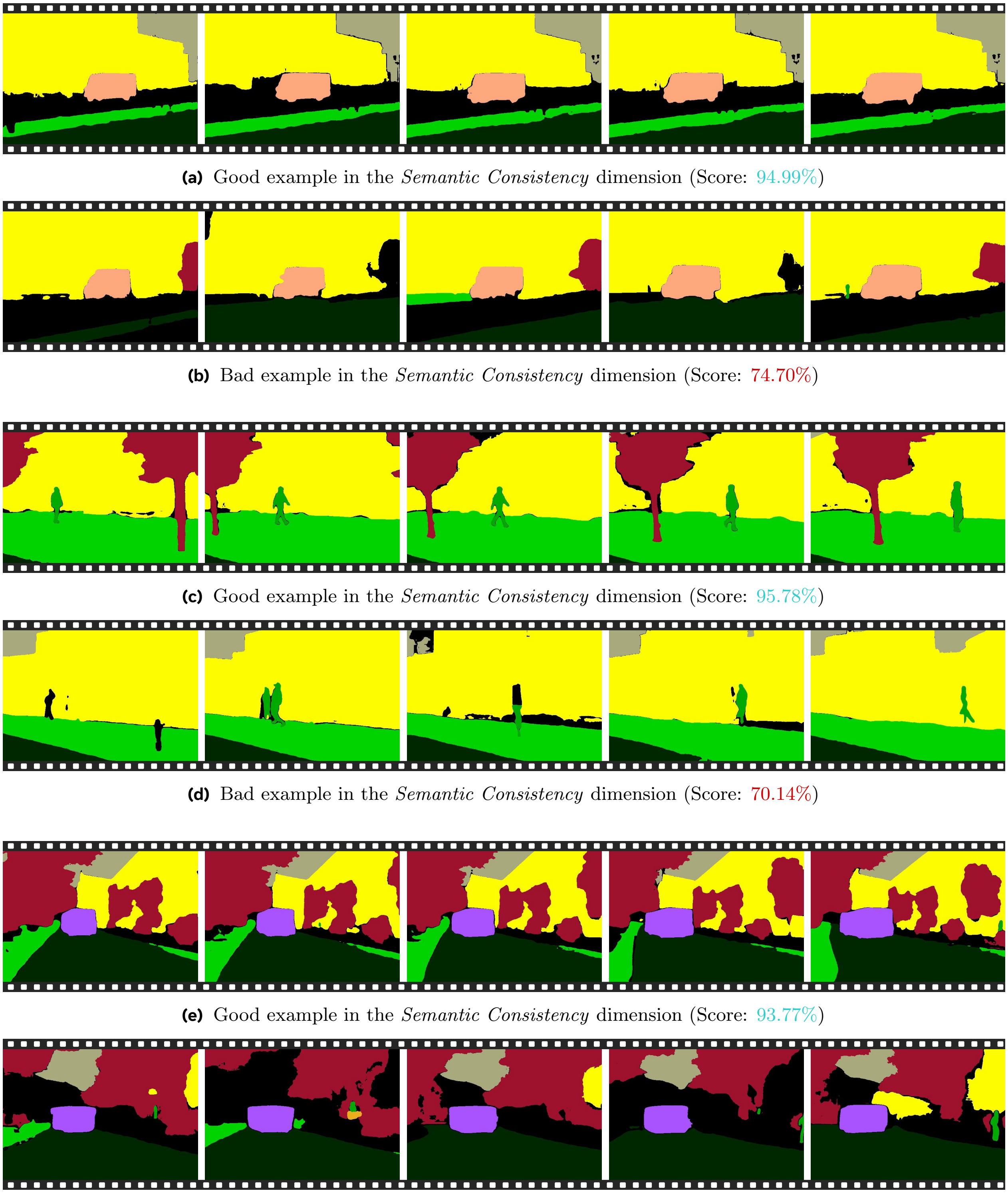
Figure 14 provides typical examples of videos with good and bad quality in terms of *Semantic Consistency*.

7.6.5 Evaluation & Analysis

Table 9 provides the complete results of models in terms of *Semantic Consistency*.

Table 9 Complete comparisons of state-of-the-art driving world models in terms of *Semantic Consistency* in WorldLens.

$\mathcal{S}_{\text{SemC}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
Label Flip Rate (LFR, ↑)	85.48%	89.15%	89.59%	88.09%	88.46%	87.92%	90.39%
Segmentation Association (SAC, ↑)	75.57%	80.85%	82.21%	79.94%	80.13%	79.54%	82.48%
Distribution Stability (CDS, ↑)	96.40%	97.31%	97.05%	96.86%	97.00%	96.95%	97.89%
Total (↑)	80.63%	84.99%	85.91%	84.08%	84.32%	83.80%	86.39%



(a) Good example in the *Semantic Consistency* dimension (Score: 94.99%)

(b) Bad example in the *Semantic Consistency* dimension (Score: 74.70%)

(c) Good example in the *Semantic Consistency* dimension (Score: 95.78%)

(d) Bad example in the *Semantic Consistency* dimension (Score: 70.14%)

(e) Good example in the *Semantic Consistency* dimension (Score: 93.77%)

(f) Bad example in the *Semantic Consistency* dimension (Score: 61.82%)

Figure 14 Examples of "good" and "bad" generation qualities in terms of *Semantic Consistency* in WorldLens.

7.7 Perceptual Discrepancy

7.7.1 Definition

Perceptual Discrepancy evaluates how closely the distribution of generated videos matches that of real ones in a learned video, semantic feature space, typically extracted by a pretrained I3D network [95] trained on Kinetics [48].

This metric captures both appearance realism and short-range temporal dynamics beyond framewise image-based metrics (*e.g.*, FID), thus reflecting the overall perceptual quality of the synthesized sequences. It is reported as a single scalar, where a lower score indicates higher perceptual similarity to real videos.

7.7.2 Formulation

Let the real and generated video sets be $\mathcal{X} = \{x_i\}_{i=1}^{N_r}$ and $\mathcal{Y} = \{y_j\}_{j=1}^{N_g}$. Each video is encoded into a d-dimensional feature vector using a fixed video encoder ϕ_{PD}:

$$\mathbf{f}_i = \phi_{\text{PD}}(x_i), \qquad \mathbf{g}_j = \phi_{\text{PD}}(y_j).$$

Let $(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ and $(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ be the empirical means and covariances of the feature sets $\{\mathbf{f}_i\}$ and $\{\mathbf{g}_j\}$, respectively. Following the Fréchet formulation, the Perceptual Fidelity score (equivalent to the Fréchet Video Distance, FVD) is defined as follows:

$$\boxed{\mathcal{S}_{\text{PD}}(\mathcal{X}, \mathcal{Y}) = \|\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\|_2^2 + \text{Tr}\Big(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y - 2(\boldsymbol{\Sigma}_x^{1/2}\boldsymbol{\Sigma}_y\boldsymbol{\Sigma}_x^{1/2})^{1/2}\Big)} \tag{7}$$

A lower $\mathcal{S}_{\text{PD}}$ indicates that the generated distribution $\mathcal{Y}$ is perceptually closer to the real distribution $\mathcal{X}$.

Perceptual Discrepancy serves as a global perceptual indicator of visual and temporal realism. By comparing distributions in a semantically informed video embedding space, it evaluates not only static appearance but also dynamic motion smoothness and coherence. A low score indicates that the generative model produces sequences with authentic spatial structures, plausible dynamics, and consistent motion statistics, while a high score reveals perceptual drift or domain mismatch.

This metric thus complements fine-grained evaluations by providing an overarching measure of distributional fidelity in world-model generation.

7.7.3 Implementation Details

Perceptual Discrepancy is measured using Fréchet Video Distance (FVD). We extract video features with a pretrained I3D model [95] (Kinetics-400 [48]) following the VideoGPT [122] protocol, and compute FVD between ground-truth and generated feature distributions.

7.7.4 Examples

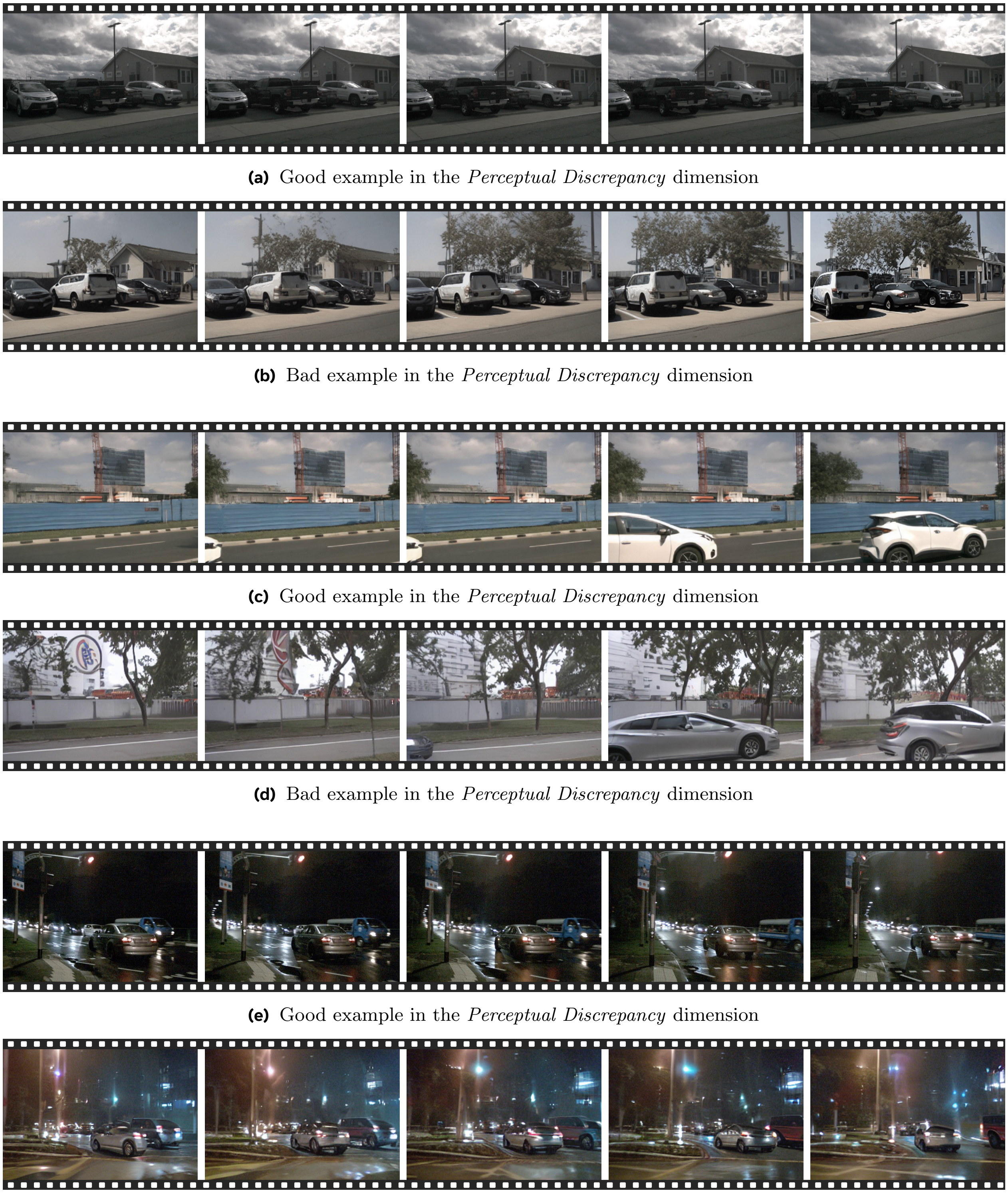
Figure 15 provides typical examples of videos with good and bad quality in terms of *Perceptual Discrepancy*.

7.7.5 Evaluation & Analysis

Table 10 provides the complete results of models in terms of *Perceptual Discrepancy*.

Table 10 Complete comparisons of state-of-the-art driving world models in terms of *Perceptual Discrepancy* in WorldLens.

$\mathcal{S}_{\text{PD}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
Total (↓)	222.00	189.76	127.07	90.42	58.08	179.74	–



(a) Good example in the *Perceptual Discrepancy* dimension

(b) Bad example in the *Perceptual Discrepancy* dimension

(c) Good example in the *Perceptual Discrepancy* dimension

(d) Bad example in the *Perceptual Discrepancy* dimension

(e) Good example in the *Perceptual Discrepancy* dimension

(f) Bad example in the *Perceptual Discrepancy* dimension

Figure 15 Examples of "good" and "bad" generation qualities in terms of *Perceptual Discrepancy* in WorldLens.

7.8 Cross-View Consistency

7.8.1 Definition

Cross-View Consistency evaluates the geometric and photometric coherence across overlapping regions between adjacent camera views in a multi-view driving scene. A spatially consistent generation should ensure that content observed from different cameras remains structurally aligned and visually coherent, faithfully representing the same physical world from multiple perspectives. This property is critical for autonomous driving, as consistent multi-view generation reflects an accurate understanding of shared 3D geometry and scene semantics.

We quantify this consistency by computing the mean accumulated confidence of feature correspondences between overlapping edge regions of adjacent camera pairs using a pretrained local feature matcher. Higher confidence indicates better geometric and appearance alignment across views.

7.8.2 Formulation

For each generated scene $y_j \in \mathcal{Y}$ with N_v synchronized views and T frames, a frozen LoFTR matcher ψ_{LoFTR} produces $M_{ab}^{(t)}$ correspondences with confidence $c_m^{(t)} \in [0,1]$ between every adjacent camera pair $(a,b) \in \mathcal{P}$ at frame t. The overall Cross-View Consistency score averages all confidences across pairs, frames, and videos:

$$\boxed{\mathcal{S}_{\text{CVC}}(\mathcal{Y}) = \frac{1}{N_g|\mathcal{P}|T} \sum_{j=1}^{N_g} \sum_{(a,b)\in\mathcal{P}} \sum_{t=1}^{T} \sum_{m=1}^{M_{ab}^{(t)}} c_m^{(t)}} \tag{8}$$

Higher $\mathcal{S}_{\text{CVC}}$ indicates stronger geometric and appearance alignment between adjacent camera views.

A high Cross-View Consistency score signifies that the generated multi-view scene maintains coherent 3D geometry and visual appearance across cameras, implying stable spatial reasoning and accurate scene composition. Conversely, low scores reveal misalignments such as perspective drift, inconsistent object boundaries, or mismatched illumination across views.

This metric thus serves as a key indicator of multi-camera integrity, linking the generative model's visual realism to its geometric understanding of the physical world.

7.8.3 Implementation Details

The *Cross-View Consistency* score is computed by extracting frame-wise sparse correspondences using the pretrained LoFTR local feature matcher [92]. Matched keypoints across views are used to assess geometric alignment between generated and ground-truth videos.

7.8.4 Examples

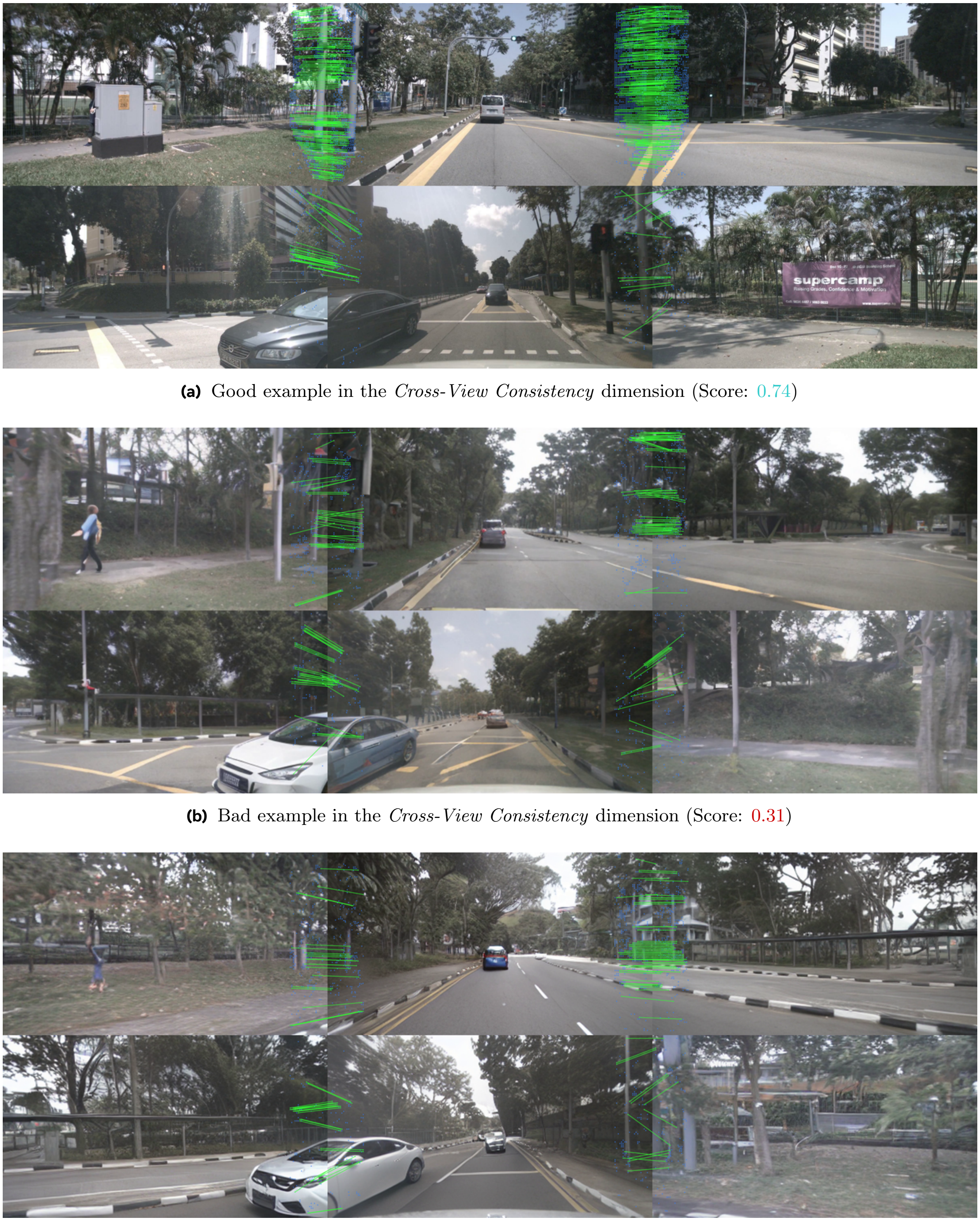
Figure 16 provides typical examples of videos with good and bad quality in terms of *Cross-View Consistency*.

7.8.5 Evaluation & Analysis

Table 11 provides the complete results of models in terms of *Cross-View Consistency*.

Table 11 Complete comparisons of state-of-the-art driving world models in terms of *Cross-View Consistency* in WorldLens.

$\mathcal{S}_{\text{CVC}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
VC Score (↑)	68.23	72.08	124.51	78.81	184.65	74.97	319.73
VC Match (↑)	185.77	194.99	302.83	211.18	389.78	201.00	570.75
Total (↑)	0.3665	0.3686	0.4065	0.3720	0.4574	0.3721	0.5420



(a) Good example in the *Cross-View Consistency* dimension (Score: 0.74)

(b) Bad example in the *Cross-View Consistency* dimension (Score: 0.31)

(c) Bad example in the *Cross-View Consistency* dimension (Score: 0.26)

Figure 16 Examples of "good" and "bad" generation qualities in terms of *Cross-View Consistency* in WorldLens.

8 Aspect 2: Reconstruction

This aspect assesses the **reconstructability** of generated videos. Given a reconstructed neural 4D representation built from each generated video, we evaluate both its internal fidelity and its rendering performance from novel viewpoints. A high-quality generation should preserve temporally coherent geometry, appearance, and illumination that jointly support faithful 4D reconstruction. We employ differentiable 4D reconstruction to optimize scene geometry and radiance from the generated sequences, then re-render the reconstructed model under both original and unseen camera poses.

8.1 Photometric Discrepancy

8.1.1 Definition

Photometric Discrepancy quantifies how accurately the 4D scene reconstructed from a generated video can reproduce its observed frames. Each generated sequence is first converted into a neural radiance field using a differentiable pipeline based on 4D Gaussian Splatting or NeRF-based [71] video reconstruction. The reconstructed model is then re-rendered from the same camera poses as the input frames, and pixel-wise fidelity is evaluated using standard image quality metrics such as PSNR, SSIM [109], and LPIPS [140].

8.1.2 Formulation

Let $\psi_{\mathrm{REC}}(\cdot)$ denote the 4D reconstruction function that produces a radiance field $\hat{\mathcal{R}}_j$ from a generated video y_j. Rendering this field at the input camera poses yields re-rendered frames: $\hat{y}_j^{(t)} = \mathrm{Render}(\hat{\mathcal{R}}_j, \mathrm{Pose}(t))$, where $\mathrm{Pose}(t)$ is the camera pose of frame t. Photometric fidelity is measured by the mean Learned Perceptual Image Patch Similarity (LPIPS) between the reconstructed and original frames:

$$\boxed{\mathcal{S}_{\mathrm{PhoF}}(\mathcal{Y}) = \tfrac{1}{N_g T} \sum\nolimits_{j=1}^{N_g} \sum\nolimits_{t=1}^{T} \mathrm{LPIPS}\Big(\hat{y}_j^{(t)},\, y_j^{(t)}\Big)} \tag{9}$$

Higher $\mathcal{S}_{\mathrm{PhoF}}$ indicates that the reconstructed radiance fields preserve fine-grained appearance details consistent with the generated frames.

8.1.3 Implementation Details

We follow the OmniRe [17] preprocessing pipeline and default configuration on nuScenes [9], using the same 6-camera setup. Each generated clip is treated as a short multi-view sequence (12 Hz, 16 frames per camera at 544×304 resolution). For each clip, we optimize a single 4D Gaussian field for 30k steps, adopting OmniRe's static- and dynamic-node Gaussian initializations [17] as well as its batch size, ray-sampling strategy, loss weights, and learning-rate schedule. After training, we render all training views and evaluate PSNR, SSIM, and LPIPS averaged over all frames and cameras.

8.1.4 Examples

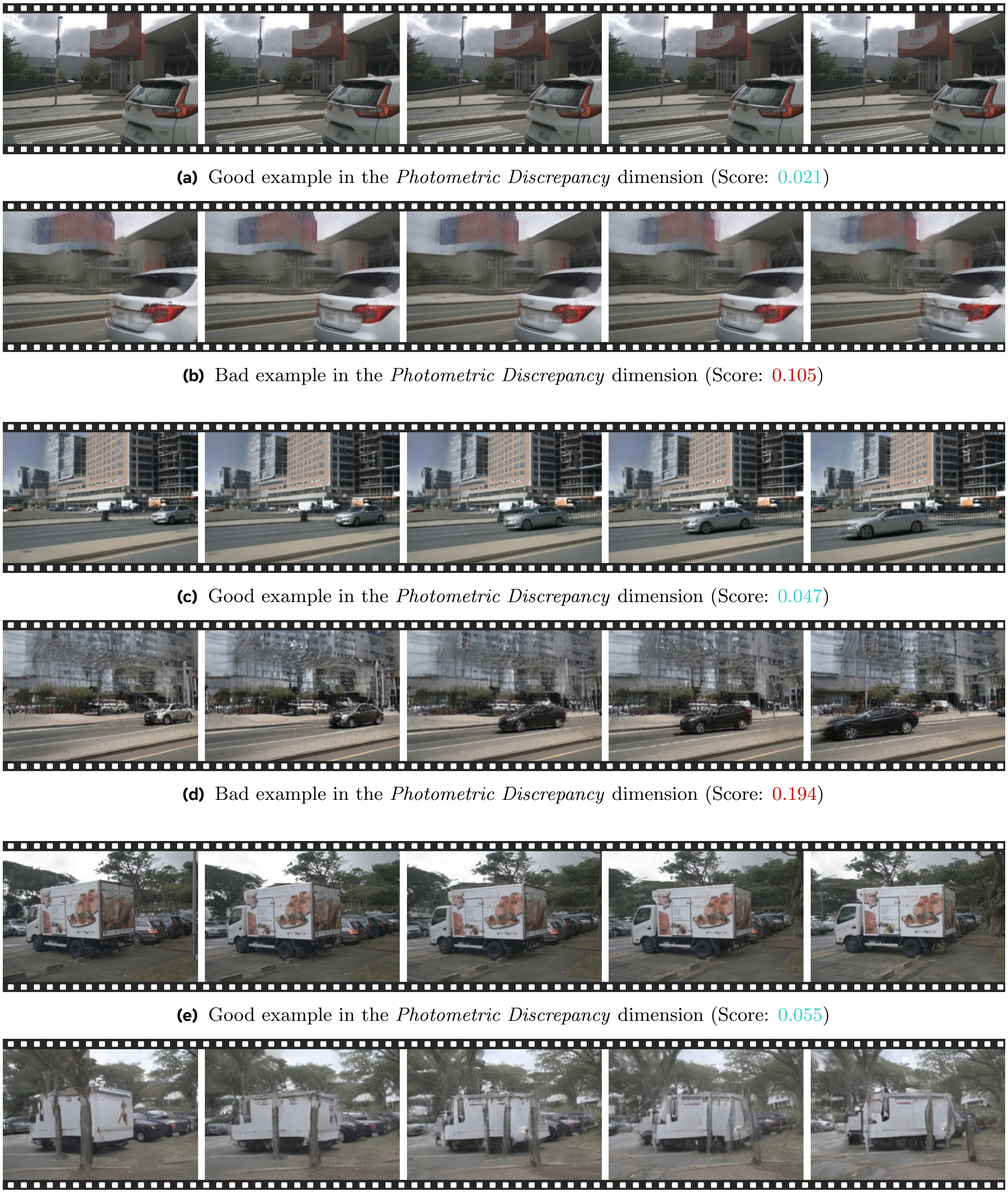
Figure 17 provides typical examples of videos with good and bad quality in terms of *Photometric Discrepancy*.

8.1.5 Evaluation & Analysis

Table 12 provides the complete results of models in terms of *Photometric Discrepancy*.

Table 12 Complete comparisons of state-of-the-art driving world models in terms of *Photometric Discrepancy* in WorldLens.

$\mathcal{S}_{\mathrm{PhoF}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
PSNR (↑)	28.44	29.11	33.15	33.21	32.89	31.25	34.31
SSIM (↑)	0.887	0.917	0.946	0.950	0.948	0.926	-
LPIPS (↓)	0.140	0.097	0.093	0.065	0.066	0.098	0.056



(a) Good example in the *Photometric Discrepancy* dimension (Score: 0.021)

(b) Bad example in the *Photometric Discrepancy* dimension (Score: 0.105)

(c) Good example in the *Photometric Discrepancy* dimension (Score: 0.047)

(d) Bad example in the *Photometric Discrepancy* dimension (Score: 0.194)

(e) Good example in the *Photometric Discrepancy* dimension (Score: 0.055)

(f) Bad example in the *Photometric Discrepancy* dimension (Score: 0.123)

Figure 17 Examples of "good" and "bad" reconstruction qualities in terms of *Photometric Discrepancy* in WorldLens.

8.2 Geometric Discrepancy

8.2.1 Definition

Geometric Discrepancy evaluates how faithfully the geometry encoded in a generated video can be recovered after reconstruction. For each generated video and its paired ground truth, we reconstruct two 4DGS models using identical camera poses and training parameters, then render per-frame depth maps for both reconstructions. Depth consistency is measured using the Absolute Relative Error (Abs Rel) computed on regions defined by Grounded-SAM 2 [80] masks that isolate road surfaces and foreground objects.

8.2.2 Formulation

Let $\psi_{\mathrm{REC}}(\cdot)$ denote the 4D reconstruction function. For each generated video y_j and ground truth x_j, we obtain two reconstructed fields $\hat{\mathcal{R}}_j = \psi_{\mathrm{REC}}(y_j)$ and $\hat{\mathcal{R}}_j^{\mathrm{gt}} = \psi_{\mathrm{REC}}(x_j)$. At each training pose $\mathrm{Pose}(t)$, the corresponding depth maps are rendered as:

$$\hat{D}_j^{(t)} = \mathrm{RenderDepth}\Big(\hat{\mathcal{R}}_j, \mathrm{Pose}(t)\Big)\,, \qquad D_j^{\mathrm{gt}(t)} = \mathrm{RenderDepth}\Big(\hat{\mathcal{R}}_j^{\mathrm{gt}}, \mathrm{Pose}(t)\Big)\,.$$

Let $M_j^{(t)}$ be the Grounded-SAM 2 mask selecting conditioned pixels. The overall Geometric Accuracy score averages the masked AbsRel error over all frames and videos:

$$\boxed{\mathcal{S}_{\mathrm{GeoA}}(\mathcal{Y}) = \frac{1}{N_g T} \sum\nolimits_{j=1}^{N_g} \sum\nolimits_{t=1}^{T} \frac{1}{|\mathcal{M}_j^{(t)}|} \sum\nolimits_{\mathbf{p}\in M_j^{(t)}} \frac{\left|\hat{D}_j^{(t)}(\mathbf{p}) - D_j^{\mathrm{gt}(t)}(\mathbf{p})\right|}{D_j^{\mathrm{gt}(t)}(\mathbf{p})}} \tag{10}$$

Lower $\mathcal{S}_{\mathrm{GeoA}}$ indicates that the reconstructed geometry from generated videos is more consistent with the ground-truth scene structure.

8.2.3 Implementation Details

This aspect shares the same training setup as the photometric discrepancy. The main difference is in the rendering and metric computation. For each clip, we render per-pixel depth from the learned 4D Gaussian field for all training views using the default Gaussian rasterizer (GSplat [134]) as in OmniRe [17], configured in the "RGB+ED" mode that outputs both color and Euclidean depth along each camera ray. To obtain fair and semantically meaningful depth metrics, we construct evaluation masks from ground-truth images using Grounded SAM 2 [80], extracting the union of the road and vehicle regions. Depth errors, *e.g.*, Abs Rel and Root Mean Squared Error (RMSE), are then computed only within these masked pixels by comparing with the depth rendered by the GT-trained Gaussian field. We also report the threshold accuracy metrics (δ_1, δ_2, δ_3). Per-clip scores are obtained by averaging over all frames and cameras of the clip.

8.2.4 Example

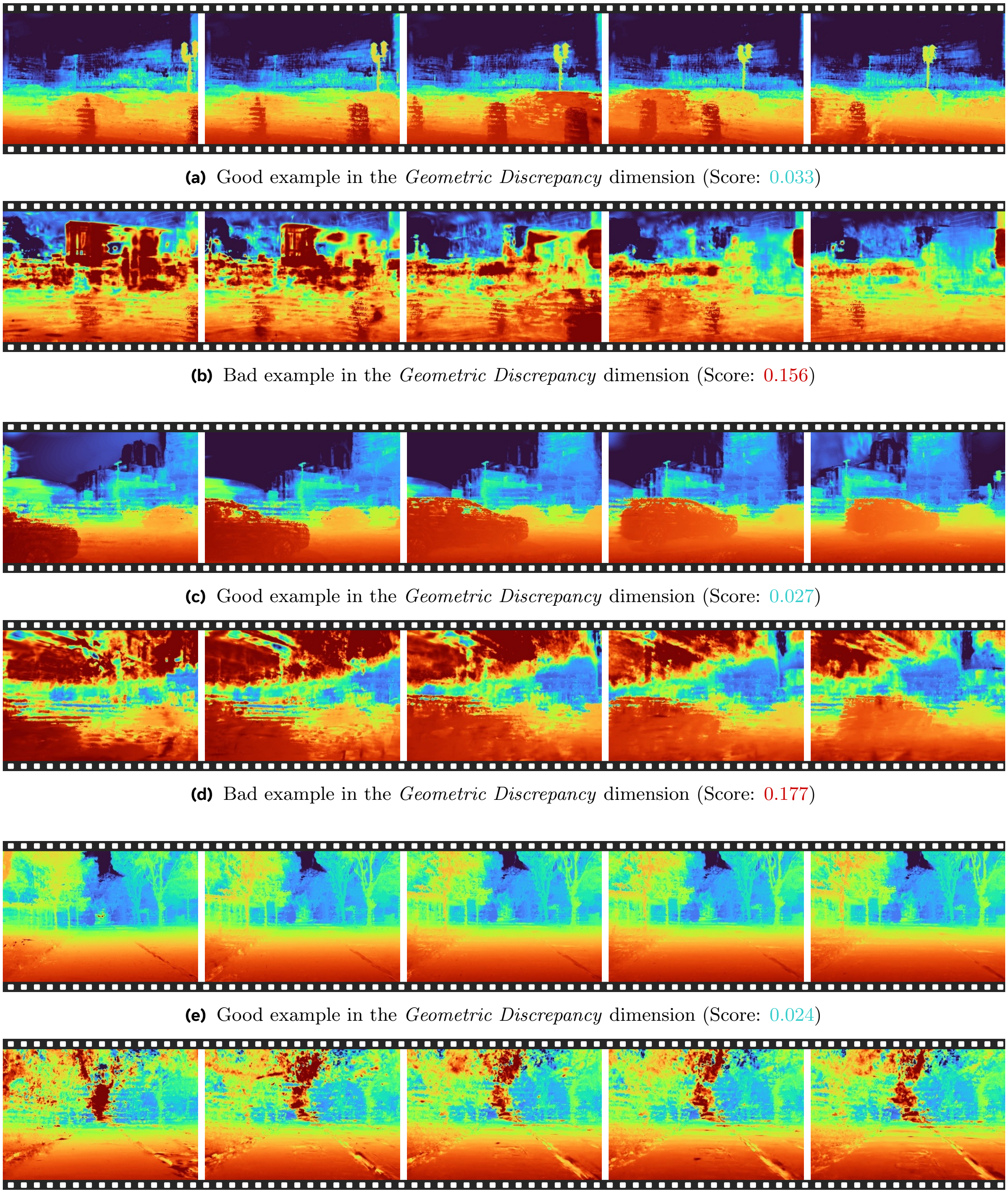
Figure 18 provides typical examples of videos with good and bad quality in terms of *Geometric Discrepancy*.

8.2.5 Evaluation & Analysis

Table 13 provides the complete results of models in terms of *Geometric Discrepancy*.

Table 13 Complete comparisons of state-of-the-art driving world models in terms of *Geometric Discrepancy* in WorldLens.

$\mathcal{S}_{\mathrm{GeoA}}(\cdot)$	**MagicDrive**	**DreamForge**	**DriveDreamer-2**	**OpenDWM**	**DiST-4D**	**$\mathcal{X}$-Scene**	**Empirical**
	[ICLR'24]	[arXiv'24]	[AAAI'25]	[CVPR'25]	[ICCV'25]	[NeurIPS'25]	**Max**
RMSE (↓)	4.116	4.166	2.869	3.130	2.969	3.594	-
Abs Rel (↓)	0.115	0.105	0.073	0.088	0.080	0.096	-
δ_1 (↑)	0.856	0.874	0.923	0.914	0.910	0.889	-
δ_2 (↑)	0.925	0.940	0.968	0.961	0.962	0.946	-
δ_3 (↑)	0.953	0.966	0.983	0.978	0.981	0.969	-



(a) Good example in the *Geometric Discrepancy* dimension (Score: 0.033)

(b) Bad example in the *Geometric Discrepancy* dimension (Score: 0.156)

(c) Good example in the *Geometric Discrepancy* dimension (Score: 0.027)

(d) Bad example in the *Geometric Discrepancy* dimension (Score: 0.177)

(e) Good example in the *Geometric Discrepancy* dimension (Score: 0.024)

(f) Bad example in the *Geometric Discrepancy* dimension (Score: 0.097)

Figure 18 Examples of "good" and "bad" reconstruction qualities in terms of *Geometric Discrepancy* in WorldLens.

8.3 Novel-View Quality

8.3.1 Definition

Novel View Quality (NVQ) assesses the perceptual quality of rendered frames from unseen camera trajectories, complementing Novel View Fidelity by focusing on frame-level realism rather than distributional similarity. For each novel-view trajectory, we render novel-view videos from reconstructed radiance fields and evaluate the perceptual quality of each frame using the pretrained MUSIQ model [49]. The novel-view trajectories are:

- "`front_center_interp`", which smoothly interpolates along the original front-center (ID 0) camera path by selecting four key poses at indices 0, $\lfloor N/4 \rfloor$, $\lfloor N/2 \rfloor$, and $\lfloor 3N/4 \rfloor$, and generating intermediate 4×4 poses through linear translation and spherical linear interpolation (Slerp) of orientations.
- "`s_curve`", which constructs an S-shaped trajectory by traversing five key poses from front-left (ID 1), front-center (ID 0), and front-right (ID 2) cameras, at indices (0), $\lfloor N/4 \rfloor$, $\lfloor N/2 \rfloor$, $\lfloor 3N/4 \rfloor$, and $(N-1)$, yielding a smooth left–center–right–center motion.
- "`lateral_offset`", which generates a parallel-view sequence by shifting each front-camera pose (ID 0) laterally by a fixed offset along its local $+x$ axis while preserving orientation, followed by temporal resampling through linear and Slerp interpolation. All trajectories are resampled to a fixed target length.

8.3.2 Formulation

Given the re-rendered novel-view videos $y_j^* = \{\text{Render}(\hat{\mathcal{R}}_j, \text{Pose}^*(t))\}_{t=1}^{T}$ under any of the novel-view settings, we compute frame-level perceptual quality scores via the pretrained image-quality assessor $\phi_{\text{MUSIQ}}(\cdot)$.

Each frame receives a quality score $q_j^{(t)} = \phi_{\text{MUSIQ}}(y_j^{*(t)})$, and the overall dataset-level Novel View Quality is obtained by averaging across all frames and videos:

$$\boxed{\mathcal{S}_{\text{NVQ}}(\mathcal{Y}) = \tfrac{1}{N_g T} \sum\nolimits_{j=1}^{N_g} \sum\nolimits_{t=1}^{T} q_j^{(t)}} \tag{11}$$

Higher $\mathcal{S}_{\text{NVQ}}$ indicates better perceptual quality of novel-view renderings, reflecting sharper appearance, fewer artifacts, and more realistic content across unseen trajectories.

8.3.3 Implementation Details

We render videos from four novel viewpoints using the Gaussian Fields trained by each world model, following the definition provided in Section 8.3.1, where the lateral offset is set to 1 m. Novel-view image quality is assessed using the pretrained MUSIQ model [49]. Each rendered novel-view video is processed frame-by-frame (resized to a maximum spatial dimension of 512 pixels), and the MUSIQ scores are averaged across all frames and videos within each view condition.

8.3.4 Examples

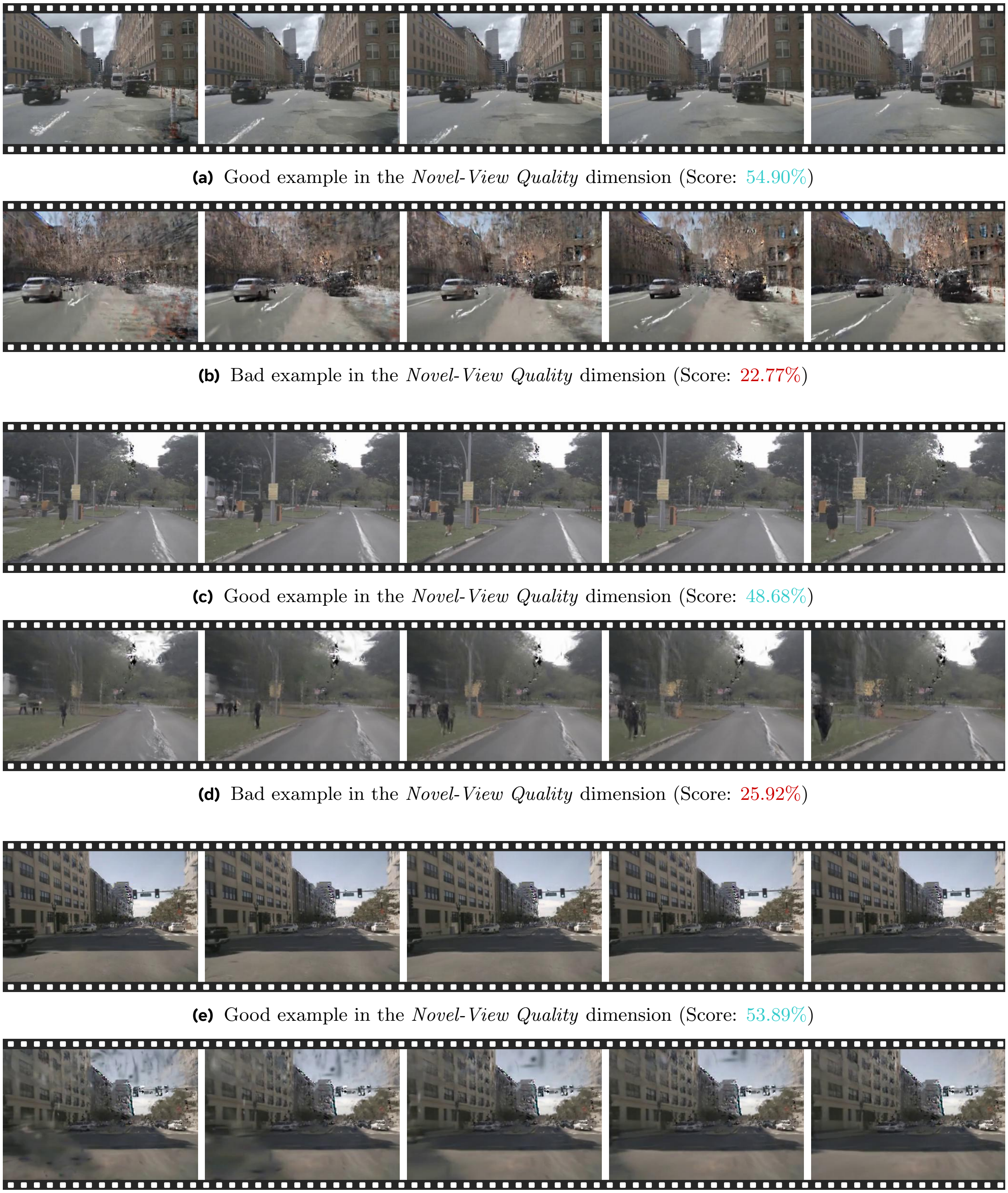
Figure 19 provides typical examples of videos with good and bad quality in terms of *Novel-View Quality.*

8.3.5 Evaluation & Analysis

Table 14 provides the complete results of models in terms of *Novel-View Quality.*

Table 14 Complete comparisons of state-of-the-art driving world models in terms of *Novel-View Quality* in WorldLens.

$\mathcal{S}_{\text{NVQ}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical** **Max**
Center Interpolation (↑)	39.31%	42.66%	37.40%	40.71%	44.67%	38.11%	-
S-Curve (↑)	39.67%	41.57%	35.15%	38.99%	42.64%	38.07%	-
Left Lateral Offset (↑)	40.28%	40.56%	36.03%	39.20%	42.64%	38.25%	-
Right Lateral Offset (↑)	40.02%	40.14%	35.82%	39.24%	42.42%	37.74%	-
Average (↑)	39.82%	41.23%	36.10%	39.54%	43.09%	38.04%	-



(a) Good example in the *Novel-View Quality* dimension (Score: 54.90%)

(b) Bad example in the *Novel-View Quality* dimension (Score: 22.77%)

(c) Good example in the *Novel-View Quality* dimension (Score: 48.68%)

(d) Bad example in the *Novel-View Quality* dimension (Score: 25.92%)

(e) Good example in the *Novel-View Quality* dimension (Score: 53.89%)

(f) Bad example in the *Novel-View Quality* dimension (Score: 28.12%)

Figure 19 Examples of "good" and "bad" reconstruction qualities in terms of *Novel-View Quality* in WorldLens.

8.4 Novel-View Discrepancy

8.4.1 Definition

Novel-View Discrepancy measures the perceptual realism of newly rendered videos under unseen camera trajectories reconstructed from generated scenes.

Given the reconstructed neural radiance field of each generated video, we render novel-view sequences at held-out camera poses and compare them against ground-truth novel-view renderings of the corresponding real scenes.

8.4.2 Formulation

Let $\hat{\mathcal{R}}_j$ and $\hat{\mathcal{R}}_j^{\text{gt}}$ denote the reconstructed radiance fields from the generated and ground-truth videos, respectively. Rendering each field along a novel trajectory $\{\text{Pose}^*(t)\}_{t=1}^{T}$ yields two new video sequences:

$$y_j^* = \{\text{Render}(\hat{\mathcal{R}}_j,\, \text{Pose}^*(t))\}_{t=1}^{T}\ , \tag{12}$$

$$x_j^* = \{\text{Render}(\hat{\mathcal{R}}_j^{\text{gt}},\, \text{Pose}^*(t))\}_{t=1}^{T}\ . \tag{13}$$

We compute the Fréchet Video Distance (FVD) between the distributions of generated and ground-truth novel-view videos using Eq. (7), where the feature extractor ϕ_{PF} (I3D on Kinetics) remains the same.

The dataset-level *Novel View Fidelity* is thus defined as:

$$\boxed{\mathcal{S}_{\text{NVD}}(\mathcal{Y}) = \mathcal{S}_{\text{PF}}\big(\{x_j^*\}, \{y_j^*\}\big)} \tag{14}$$

Lower $\mathcal{S}_{\text{NVD}}$ indicates higher perceptual fidelity of the reconstructed scenes when viewed from unseen camera trajectories.

8.4.3 Implementation Details

The selection of novel viewpoints and the rendering configurations are kept consistent with those in Section 8.3. For *Novel-View Discrepancy*, the calculation process follows the same setting of Section 7.7.

8.4.4 Example

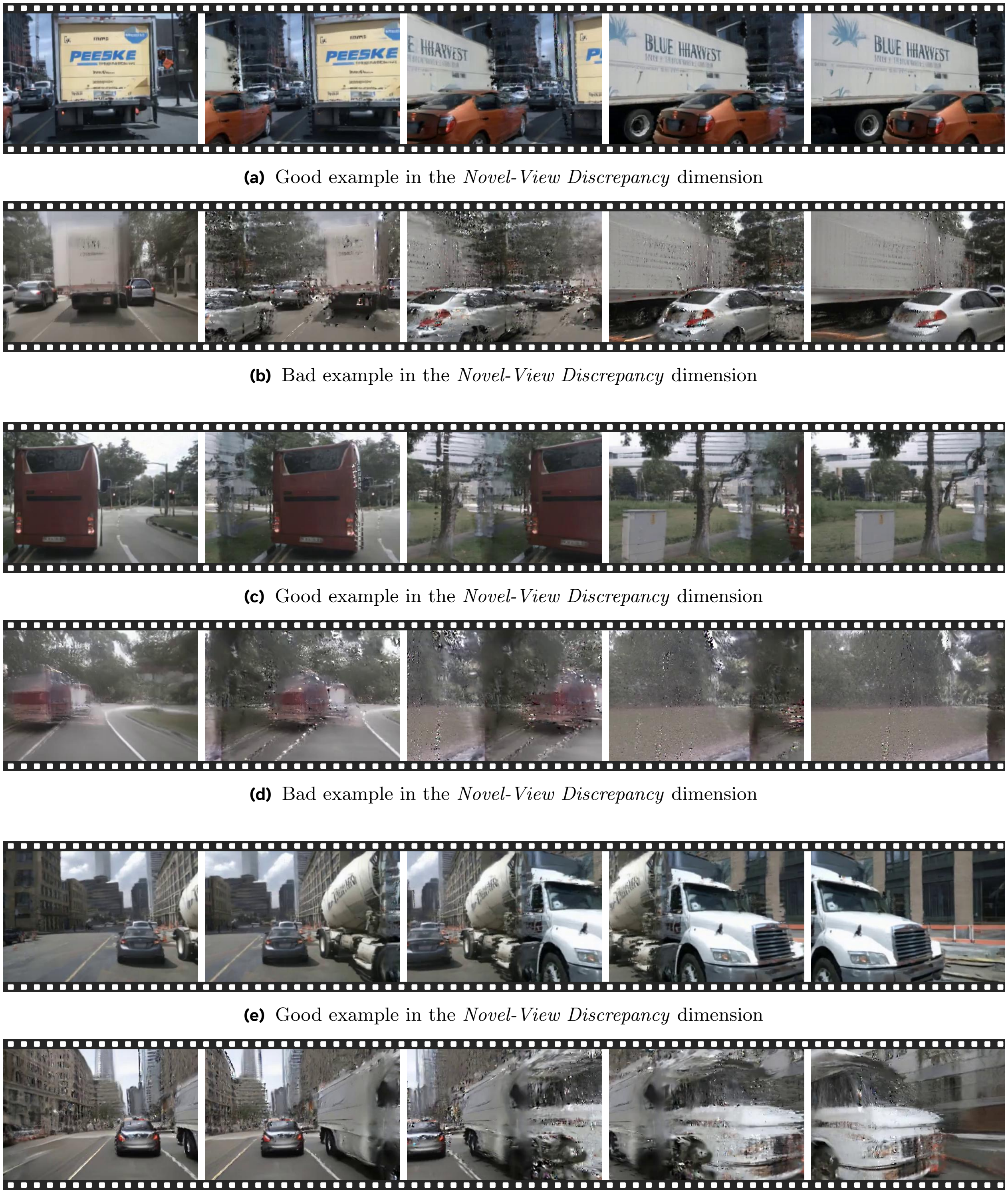
Figure 20 provides typical examples of videos with good and bad quality in terms of *Novel-View Discrepancy*.

8.4.5 Evaluation & Analysis

Table 15 provides the complete results of models in terms of *Novel-View Discrepancy*.

Table 15 Complete comparisons of state-of-the-art driving world models in terms of *Novel-View Discrepancy* in WorldLens.

$\mathcal{S}_{\text{NVD}}(\cdot)$	**MagicDrive**	**DreamForge**	**DriveDreamer-2**	**OpenDWM**	**DiST-4D**	**$\mathcal{X}$-Scene**	**Empirical**
	[ICLR'24]	[arXiv'24]	[AAAI'25]	[CVPR'25]	[ICCV'25]	[NeurIPS'25]	**Max**
Center Interpolation (↓)	448.62	403.47	259.96	339.85	190.17	376.67	-
S-Curve (↓)	281.91	171.93	132.68	159.84	96.90	219.89	-
Left Lateral Offset (↓)	492.97	400.68	326.95	318.26	237.08	435.50	-
Right Lateral Offset (↓)	485.70	414.72	320.05	332.97	245.42	430.78	-
Average (↓)	427.30	347.70	259.91	287.73	192.39	365.71	-



(a) Good example in the *Novel-View Discrepancy* dimension

(b) Bad example in the *Novel-View Discrepancy* dimension

(c) Good example in the *Novel-View Discrepancy* dimension

(d) Bad example in the *Novel-View Discrepancy* dimension

(e) Good example in the *Novel-View Discrepancy* dimension

(f) Bad example in the *Novel-View Discrepancy* dimension

Figure 20 Examples of "good" and "bad" reconstruction qualities in terms of *Novel-View Discrepancy* in WorldLens.

9 Aspect 3: Action-Following

In this section, we evaluate the **Action-Following** capability of driving world models, which reflects how well the generated videos preserve the functional cues necessary for downstream decision-making and control. Here, we assess the *functional alignment* between generated content and real-world driving behavior. Specifically, we examine how the visual information synthesized influences an end-to-end planning agent in both **open-loop** and **closed-loop** simulation settings. A model with strong action-following ability should not only generate visually convincing scenes but also guide a pretrained planner to produce trajectories and control actions that are consistent with those derived from real-world videos.

9.1 Displacement Error

9.1.1 Definition

Displacement Error (L2) evaluates the functional consistency of generated videos on the downstream task of motion planning. Instead of measuring perceptual realism or pixel-level accuracy, this metric assesses whether a generated video can serve as a reliable input for an end-to-end planner. It measures how closely the predicted trajectory inferred from a generated video aligns with the trajectory predicted from the corresponding ground-truth video. A lower displacement error indicates that the generated sequence preserves the semantic and motion cues necessary for robust trajectory forecasting, demonstrating that it is not only visually plausible but also functionally faithful to real-world driving dynamics.

9.1.2 Formulation

We employ a pretrained end-to-end planning network $\psi_{\text{Plan}}(\cdot)$ to predict trajectories from both generated and ground-truth videos. Given paired sequences y_j and x_j, the model produces corresponding planned trajectories

$$\hat{\tau}_j^{\text{gen}} = \psi_{\text{Plan}}(y_j), \qquad \hat{\tau}_j^{\text{gt}} = \psi_{\text{Plan}}(x_j),$$

where each trajectory $\hat{\tau} \in \mathbb{R}^{T_p \times 2}$ contains T_p future waypoints in 2D ground-plane coordinates. The Displacement Error is computed as the mean L2 distance between corresponding waypoints:

$$\boxed{\mathcal{S}_{\text{DE}}(\mathcal{Y}) = \frac{1}{N_g T_p} \sum_{j=1}^{N_g} \sum_{t=1}^{T_p} \left\| \hat{\tau}_j^{\text{gen}}(t) - \hat{\tau}_j^{\text{gt}}(t) \right\|_2} \tag{15}$$

Lower $\mathcal{S}_{\text{DE}}$ indicates that the generated videos induce planning behaviors that are more consistent with those derived from real-world observations, reflecting higher functional fidelity.

9.1.3 Implementation Details

We conduct the *Displacement Error* evaluation on the official nuScenes validation set, which consists of 150 diverse driving scenes. For each test case, the driving world model generates a video sequence conditioned on the initial context. These synthesized videos are then used as input for UniAD [38], a state-of-the-art end-to-end planning network, to infer future ego-motion trajectories. Following the standard protocol, we extract the planned trajectory for a horizon of 1 second (covering the immediate future waypoints). The *Displacement Error* is calculated as the L2 distance between the trajectory predicted from the generated video and the trajectory predicted from the ground-truth video. This metric strictly isolates the impact of visual generation quality on downstream perception and planning accuracy in an open-loop setting.

9.1.4 Examples

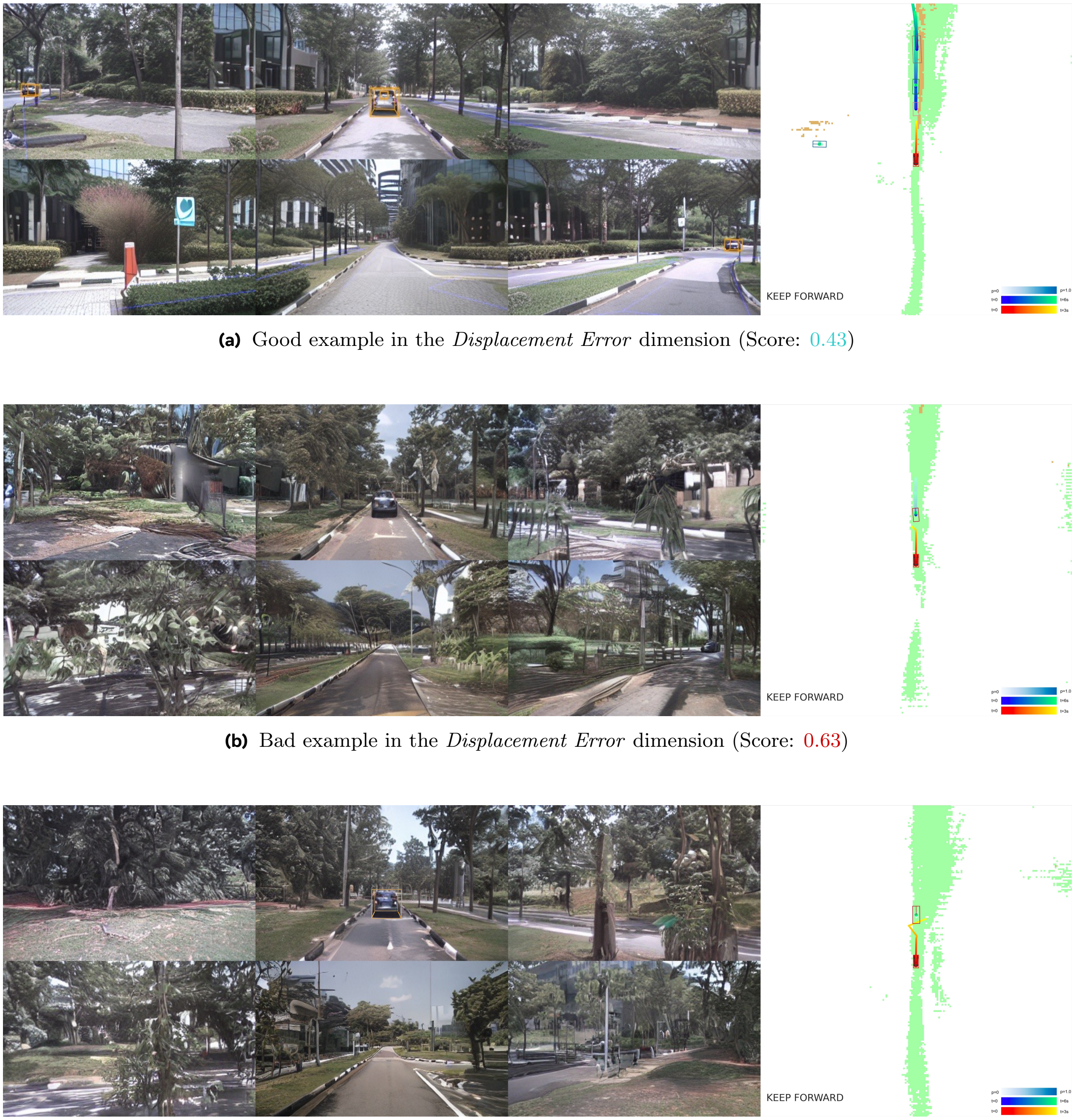
Figure 21 provides typical examples of videos with good and bad quality in terms of *Displacement Error*.

9.1.5 Evaluation & Analysis

Table 16 provides the complete results of models in terms of *Displacement Error*.

Table 16 Complete comparisons of state-of-the-art driving world models in terms of *Displacement Error* in WorldLens.

$\mathcal{S}_{\text{DE}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**Panacea** [CVPR'24]	**DrivingSphere** [CVPR'25]	**MagicDrive-V2** [ICCV'25]	**RLGF** [NeurIPS'25]	**Empirical Max**
Total (↓)	0.57	0.57	0.58	0.55	0.54	0.53	0.51



(a) Good example in the *Displacement Error* dimension (Score: 0.43)

(b) Bad example in the *Displacement Error* dimension (Score: 0.63)

(c) Bad example in the *Displacement Error* dimension (Score: 0.71)

Figure 21 Examples of "good" and "bad" action-following performances in terms of *Displacement Error* in WorldLens.

9.2 Open-Loop Adherence

9.2.1 Definition

Open-Loop Adherence evaluates the functional reliability of generated videos by measuring how well an end-to-end driving policy can perform when operating on the generated input in a non-reactive simulation environment. Following NAVSIM [20], we use the *Predictive Driver Model Score (PDMS)* to quantify adherence between the policy behavior induced by generated videos and that observed under real data.

9.2.2 Formulation

Given a pretrained planner $\psi_{\text{Plan}}(\cdot)$ and its predicted trajectory $\hat{\tau}_j$ from a generated video y_j, we simulate the resulting ego motion over a fixed horizon (*e.g.*, 4 s) in a non-reactive setting where other agents follow their recorded trajectories. At each timestep, sub-scores are computed for: *no collision* (NC), *drivable-area compliance* (DAC), *ego progress* (EP), *time-to-collision* (TTC), and *comfort* (C). Penalties (NC, DAC) suppress inadmissible behaviors, while the remaining terms are averaged with fixed weights. The PDMS is defined as:

$$\text{PDMS} = \Big(\prod_{m\in\{\text{NC},\text{DAC}\}} \text{score}_m \Big) \cdot \frac{\sum_{w\in\{\text{EP},\text{TTC},\text{C}\}} \text{weight}_w \, \text{score}_w}{\sum_{w\in\{\text{EP},\text{TTC},\text{C}\}} \text{weight}_w} \ ,$$

with default weights $\text{weight}_{\text{EP}}$=$\text{weight}_{\text{TTC}}$=5 and weight_{C}=2 as in [20]. We report the dataset-level score as the mean PDMS across all evaluated videos:

$$\boxed{\mathcal{S}_{\text{PDMS}}(\mathcal{Y}) = \tfrac{1}{N_g} \sum_{j=1}^{N_g} \text{PDMS}(y_j)} \tag{16}$$

Higher $\mathcal{S}_{\text{PDMS}}$ indicates stronger alignment between generated and real scenes in terms of functional behavior.

9.2.3 Implementation Details

We support two map environments, *singapore-onenorth* and *boston-seaport*, aligned with the DriveArena platform [127]. A total of five simulation sequences are defined for validation, enabling the evaluation of driving agents in both open-loop and closed-loop modes. In our implementation, the traffic flow engine [111] operates at a frequency of 10 Hz, while the control signals are set to 2 Hz. Every 0.5 simulation seconds, the 2D traffic flow engine updates its state and renders multi-view layouts as conditions for the video generation model. Video generation models use the last 3 frames as reference images to generate 448 × 800 images, which are subsequently resized to 224 × 400 to serve as input for the driving agent.

9.2.4 Examples

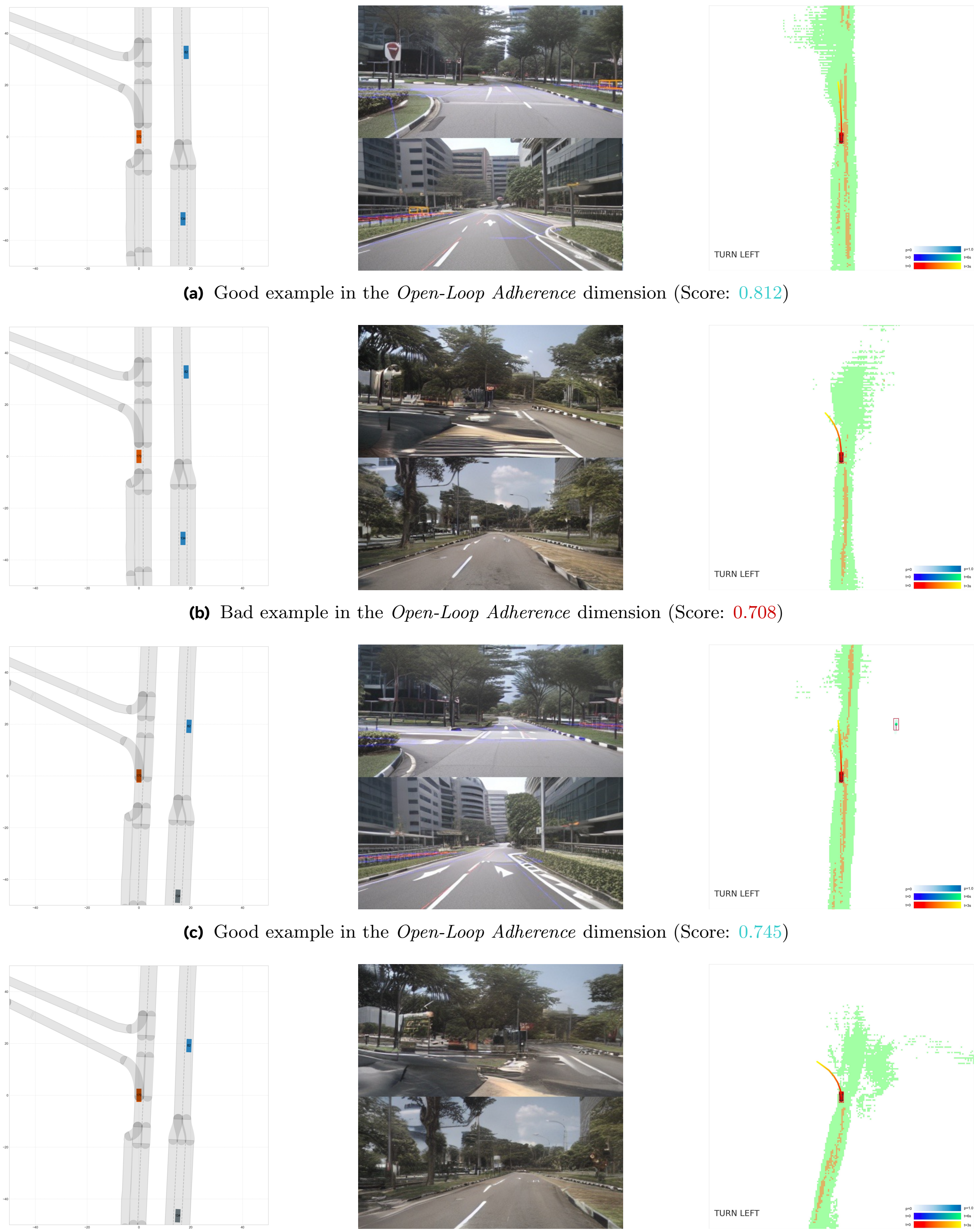
Figure 22 provides typical examples of videos with good and bad quality in terms of *Open-Loop Adherence*.

9.2.5 Evaluation & Analysis

Table 17 provides the complete results of models in terms of *Open-Loop Adherence*.

Table 17 Complete comparisons of state-of-the-art driving world models in terms of *Open-Loop Adherence* in WorldLens.

$\mathcal{S}_{\text{PDMS}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DrivingSphere** [CVPR'25]	**MagicDrive-V2** [ICCV'25]	**RLGF** [NeurIPS'25]	**Empirical** **Max**
No Collision (NC, ↑)	0.885	0.915	0.932	0.968	0.975	-
Compliance (DAC, ↑)	0.955	0.970	0.978	0.985	0.988	-
Ego Progress (EP, ↑)	0.825	0.832	0.835	0.842	0.838	-
Time-to-Collision (TTC, ↑)	0.840	0.855	0.860	0.865	0.850	-
Comfort (C, ↑)	0.815	0.825	0.830	0.835	0.828	-
Total (↑)	0.712	0.755	0.760	0.789	0.784	-



(a) Good example in the *Open-Loop Adherence* dimension (Score: 0.812)

(b) Bad example in the *Open-Loop Adherence* dimension (Score: 0.708)

(c) Good example in the *Open-Loop Adherence* dimension (Score: 0.745)

(d) Bad example in the *Open-Loop Adherence* dimension (Score: 0.621)

Figure 22 Examples of "good" and "bad" action-following performances in terms of *Open-Loop Adherence* in WorldLens.

9.3 Route Completion

9.3.1 Definition

Route Completion (RC) measures the ability of an autonomous driving agent to complete a predefined navigation route in closed-loop simulation. It quantifies the percentage of the total planned route distance successfully traveled by the ego agent before simulation termination (*e.g.*, collision, off-road, or timeout). Higher RC values indicate better long-horizon stability and control consistency, reflecting how well the generated video enables the policy to sustain safe driving behavior throughout the route.

9.3.2 Formulation

Let D_{total} denote the total length of the planned route, and $D_{\text{completed}}$ the distance actually traveled by the ego agent before termination. Following [20, 127], *Route Completion* is defined as the ratio between the completed and total distances:

$$\mathrm{RC} = \frac{D_{\text{completed}}}{D_{\text{total}}} .$$

We report the dataset-level metric as the mean RC across all evaluated closed-loop rollouts:

$$\boxed{\mathcal{S}_{\mathrm{RC}}(\mathcal{Y}) = \tfrac{1}{N_g} \sum\nolimits_{j=1}^{N_g} \mathrm{RC}(y_j)} \tag{17}$$

Higher $\mathcal{S}_{\mathrm{RC}}$ indicates that the generated scenes enable the planner to complete longer portions of the route, implying greater action stability and environmental consistency.

9.3.3 Implementation Details

Different from the open-loop evaluation (Displacement Error), both Route Completion and Closed-Loop Adherence are evaluated in a fully reactive closed-loop mode. In this setting, the ego-vehicle's trajectory is not determined by pre-recorded logs but is driven by the agent's decisions.

Specifically, the planning agent processes the video generated by the world model, outputs a control signal, and this signal updates the ego-vehicle's state within the simulator.

The world model then generates the next frame based on this new state, creating a continuous feedback loop. A simulation episode continues until one of the following termination criteria is met:

1. *Completion:* The agent successfully reaches the destination and finishes the predefined route.
2. *Failure:* The simulation is terminated early due to safety-critical infractions, specifically collision with other objects or driving off-road (exiting the drivable area).

This setup evaluates the ability of the generative driving world model to support long-horizon consistency and error-free decision-making.

9.3.4 Examples

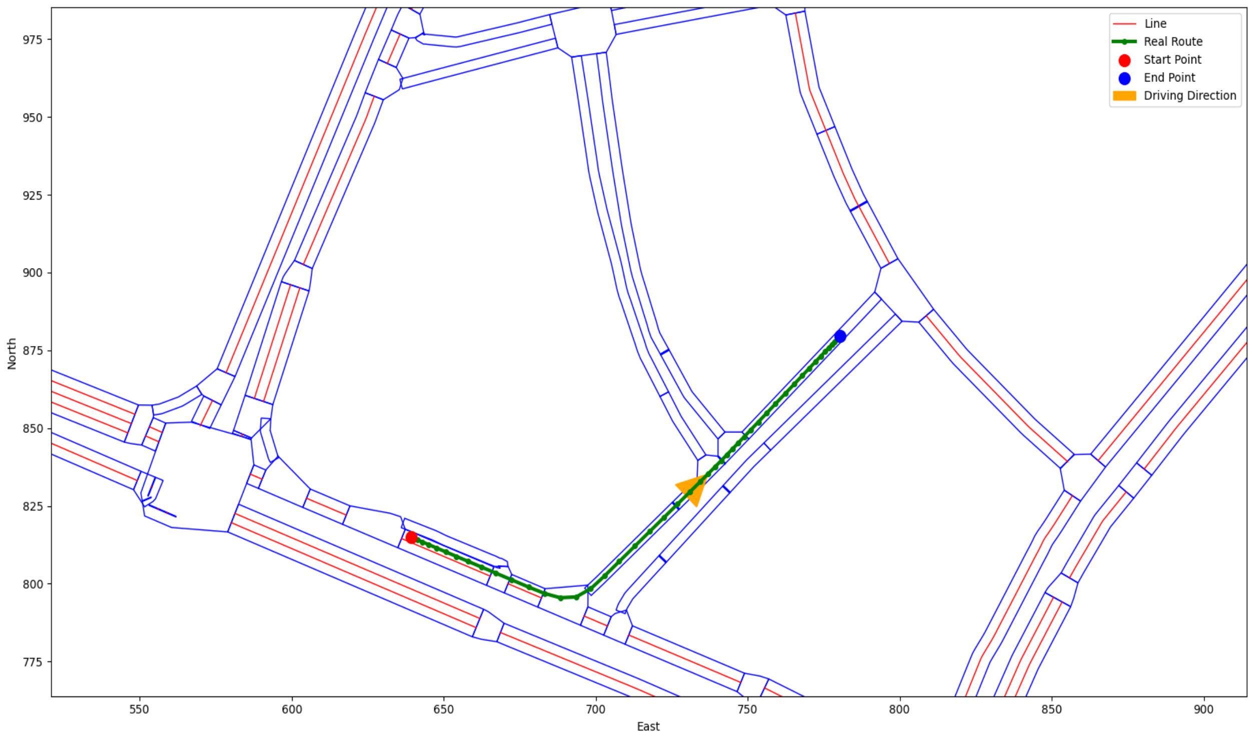
Figure 23 provides typical examples of videos with good and bad quality in terms of *Route Completion.*

9.3.5 Evaluation & Analysis

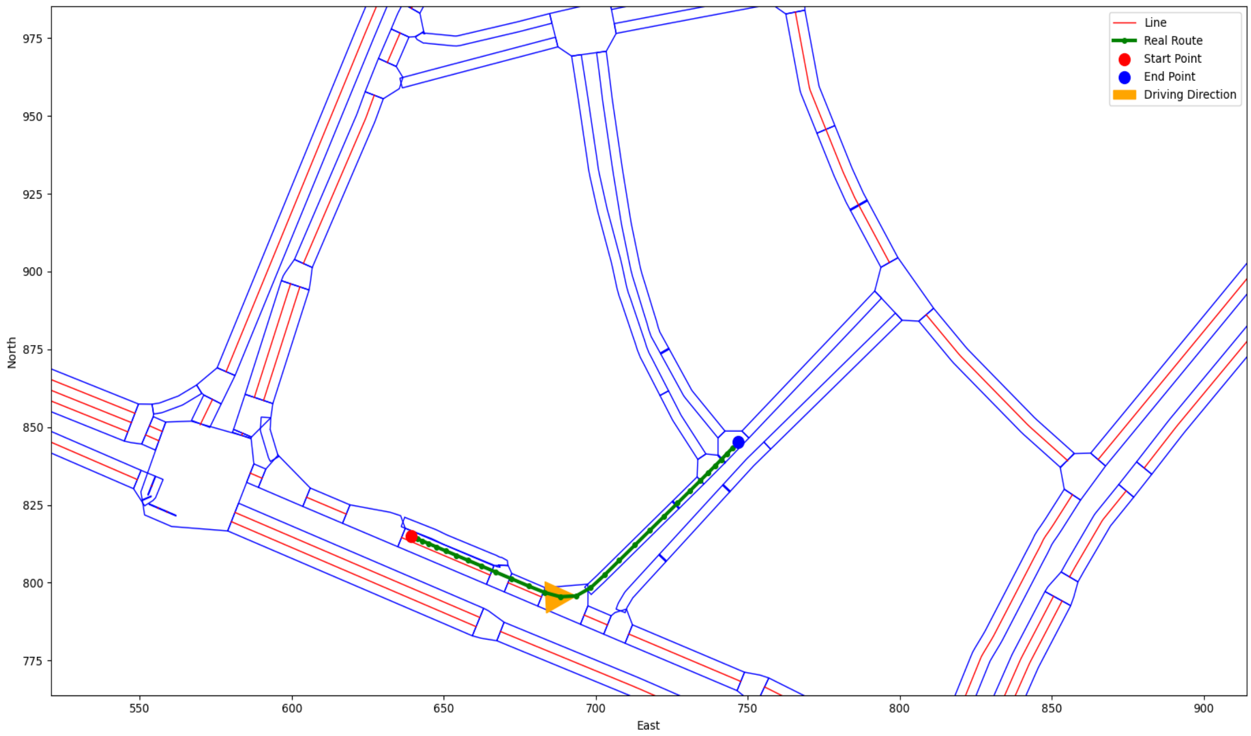
Table 18 provides the complete results of models in terms of *Route Completion.*

Table 18 Complete comparisons of state-of-the-art driving world models in terms of *Route Completion* in WorldLens

$\mathcal{S}_{\mathrm{RC}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**Panacea** [CVPR'24]	**DrivingSphere** [CVPR'25]	**MagicDrive-V2** [ICCV'25]	**RLGF** [NeurIPS'25]	**Empirical Max**
Total (↑)	6.89%	10.23%	-	11.02%	12.31%	13.51%	-



(a) Good example in the *Route Completion* dimension (Score: 18.7%)



(b) Bad example in the *Route Completion* dimension (Score: 11.2%)

Figure 23 Examples of "good" and "bad" action-following performances in terms of *Route Completion* in WorldLens.

9.4 Closed-Loop Adherence

9.4.1 Definition

Closed-Loop Adherence measures the overall driving performance of an autonomous agent in a closed-loop simulation. It is represented by the *Arena Driving Score (ADS)* [127], which jointly accounts for both driving quality and task completion.

While the PDMS score reflects the safety, comfort, and stability of the predicted trajectory, the Route Completion (RC) measures how much of the planned route is successfully finished without failure. The multiplicative formulation ensures that an agent must be both competent (high PDMS) and consistent (high RC) to achieve a strong overall score. Agents that drive perfectly but crash early, or complete the route with poor motion quality, will both be penalized accordingly.

9.4.2 Formulation

Given the PDMS and RC metrics defined in (9.2.2) and (9.3.2), the *Arena Driving Score (ADS)* is computed as follows:

$$\text{ADS} = \text{RC} \times \text{PDMS} \ ,$$

where $\text{RC} \in [0, 1]$ denotes route completion. For a dataset of generated videos $\mathcal{Y}$, the final closed-loop adherence is reported as the mean ADS across all evaluated driving episodes:

$$\boxed{\mathcal{S}_{\text{ADS}}(\mathcal{Y}) = \frac{1}{N_g} \sum_{j=1}^{N_g} \text{ADS}(y_j)} \tag{18}$$

Higher $\mathcal{S}_{\text{ADS}}$ indicates that the generated videos yield planners capable of both safe and complete driving behavior in closed-loop simulation.

9.4.3 Implementation Details

Closed-Loop Adherence shares the same experiment environment with *Route Completion*. The implementation details can be found in Section 9.3.

9.4.4 Examples

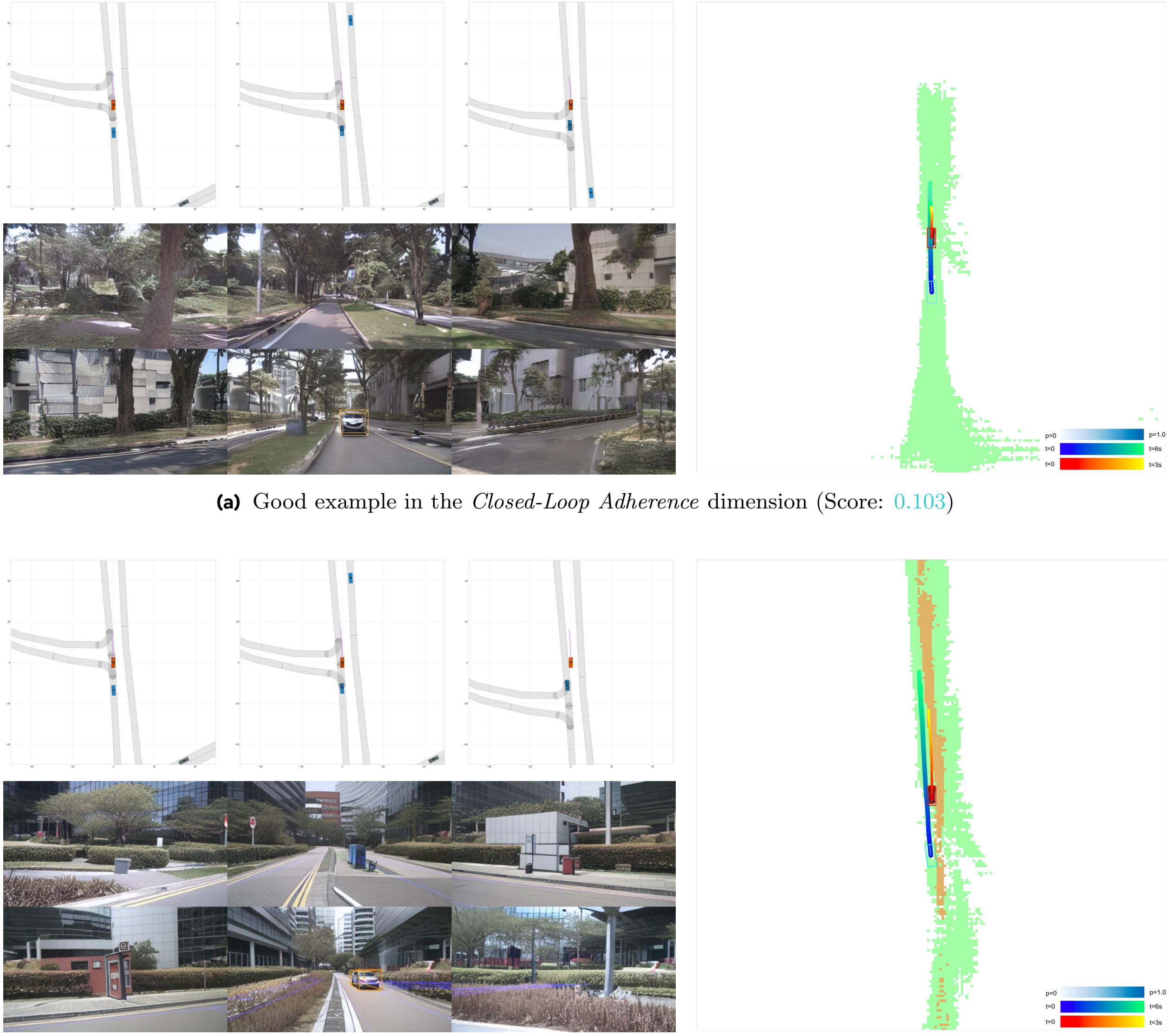
Figure 24 provides typical examples of videos with good and bad quality in terms of *Closed-Loop Adherence*.

9.4.5 Evaluation & Analysis

Table 19 provides the complete results of models in terms of *Closed-Loop Adherence*.

Table 19 Complete comparisons of state-of-the-art driving world models in terms of *Closed-Loop Adherence* in WorldLens.

$\mathcal{S}_{\text{ADS}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DrivingSphere** [CVPR'25]	**MagicDrive-V2** [ICCV'25]	**RLGF** [NeurIPS'25]	**Empirical Max**
No Collision (NC, ↑)	0.815	0.855	0.858	0.885	0.912	-
Compliance (DAC, ↑)	0.910	0.930	0.935	0.948	0.965	-
Ego Progress (EP, ↑)	0.712	0.740	0.745	0.770	0.985	-
Time-to-Collision (TTC, ↑)	0.745	0.765	0.772	0.795	0.905	-
Comfort (C, ↑)	0.720	0.745	0.750	0.765	0.850	-
Route Completion (RC, ↑)	0.068	0.102	0.110	0.123	0.135	-
Total (↑)	0.048	0.077	0.083	0.095	0.106	-



(a) Good example in the *Closed-Loop Adherence* dimension (Score: 0.103)

(b) Bad example in the *Closed-Loop Adherence* dimension (Score: 0.062)

Figure 24 Examples of "good" and "bad" action-following performances in terms of *Closed-Loop Adherence* in WorldLens.

10 Aspect 4: Downstream Task

In this section, we evaluate the **downstream task utility** of generated videos by assessing how well pretrained perception models perform when applied to synthetic data. Rather than measuring visual realism or temporal stability directly, this aspect examines whether a generative world model can produce data that is *useful* for real-world perception tasks. Specifically, we test **four representative downstream tasks** that span spatial understanding, object reasoning, and 3D scene interpretation. For each task, a perception model is pretrained on the corresponding ground-truth dataset and then evaluated on videos generated by the world model. Performance degradation relative to the ground truth reflects the distribution gap introduced by generation.

10.1 Map Segmentation

10.1.1 Definition

BEV (Bird's-Eye-View) Map Segmentation evaluates whether individual generated frames contain sufficient spatial and semantic cues for top-down mapping. A pretrained perception network $\psi_{\mathrm{BEV}}(\cdot)$ takes each generated frame $y_j^{(t)}$ as input and predicts a BEV semantic map, which is compared with the corresponding ground-truth annotation using mean Intersection-over-Union (mIoU). Higher scores indicate that the generated frames preserve structural layout and scene semantics conducive to reliable map inference.

10.1.2 Formulation

For each generated frame $y_j^{(t)}$, the pretrained model predicts a BEV map: $\hat{B}_j^{(t)} = \psi_{\mathrm{BEV}}(y_j^{(t)})$ and $\hat{B}_j^{(t)} \in \{0, \ldots, C_{\mathrm{BEV}}-1\}^{H_b \times W_b}$, and $B_j^{\mathrm{gt}(t)}$ denotes the corresponding ground-truth BEV annotation. The per-frame mean IoU is computed as:

$$S^{(t)}_{\mathrm{BEV}}(y_j) = \frac{1}{C_{\mathrm{BEV}}} \sum_{c=1}^{C_{\mathrm{BEV}}} \frac{|\hat{B}_j^{(t,c)} \cap B_j^{\mathrm{gt}(t,c)}|}{|\hat{B}_j^{(t,c)} \cup B_j^{\mathrm{gt}(t,c)}|}\ .$$

The dataset-level Map Segmentation score averages over all frames and videos, that is:

$$\boxed{\mathcal{S}_{\mathrm{Seg}}(\mathcal{Y}) = \frac{1}{N_g T} \sum_{j=1}^{N_g} \sum_{t=1}^{T} S^{(t)}_{\mathrm{BEV}}(y_j)} \tag{19}$$

where C_{BEV} is the number of BEV categories and (H_b, W_b) the BEV map resolution.

10.1.3 Implementation Details

We employ the pretrained BEVFusion multi-task model of Liu *et al.* [66], using its camera-only configuration with a ResNet-101 [33] backbone and BEVFormer encoder. The model predicts BEV semantic maps on a 150×150 grid covering a $[-30, 30] \times [-15, 15]$ m region, which are used for mIoU evaluation.

10.1.4 Examples

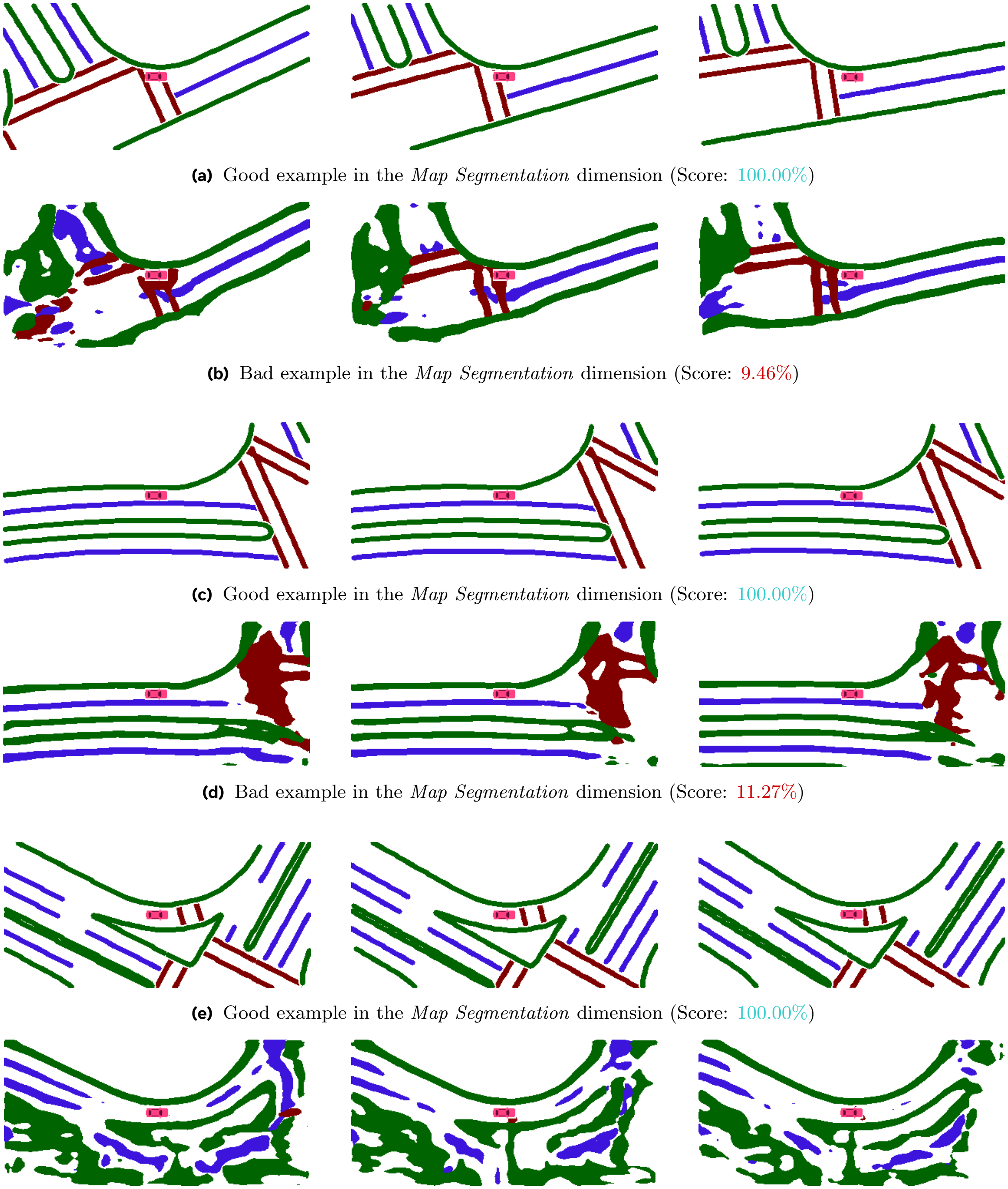
Figure 25 provides typical examples of videos with good and bad quality in terms of *Map Segmentation*.

10.1.5 Evaluation & Analysis

Table 20 provides the complete results of models in terms of *Map Segmentation*.

Table 20 Complete comparisons of state-of-the-art driving world models in terms of *Map Segmentation* in WorldLens.

$\mathcal{S}_{\mathrm{Seg}}(\cdot)$	**MagicDrive**	**DreamForge**	**DriveDreamer-2**	**OpenDWM**	**DiST-4D**	**$\mathcal{X}$-Scene**	**Empirical**
	[ICLR'24]	[arXiv'24]	[AAAI'25]	[CVPR'25]	[ICCV'25]	[NeurIPS'25]	**Max**
Divider (↑)	23.39%	34.44%	39.69%	31.88%	41.26%	31.84%	46.08%
Ped. Crossing (↑)	9.77%	21.18%	24.12%	20.27%	26.17%	17.67%	30.38%
Boundary (↑)	21.87%	35.31%	37.03%	30.74%	39.23%	32.22%	45.45%
Average (↑)	18.34%	30.31%	33.62%	27.63%	35.55%	27.24%	40.64%



(a) Good example in the *Map Segmentation* dimension (Score: 100.00%)

(b) Bad example in the *Map Segmentation* dimension (Score: 9.46%)

(c) Good example in the *Map Segmentation* dimension (Score: 100.00%)

(d) Bad example in the *Map Segmentation* dimension (Score: 11.27%)

(e) Good example in the *Map Segmentation* dimension (Score: 100.00%)

(f) Bad example in the *Map Segmentation* dimension (Score: 7.88%)

Figure 25 Examples of "good" and "bad" downstream task performances in terms of *Map Segmentation* in WorldLens.

10.2 3D Object Detection

10.2.1 Definition

3D Object Detection evaluates whether generated frames preserve the geometric and motion cues necessary for accurate perception of traffic participants.

A pretrained detector $\psi_{\mathrm{DET}}(\cdot)$, trained on ground-truth data, is applied to each generated frame $y_j^{(t)}$ to predict 3D bounding boxes with category, position, scale, and velocity information. Following the nuScenes detection protocol [9], detections are compared against ground-truth boxes to compute mean Average Precision (mAP) and the consolidated nuScenes Detection Score (NDS).

Higher mAP and NDS indicate that the generated data retains faithful 3D spatial structure and dynamic cues consistent with real-world scenes.

10.2.2 Formulation

For each frame $y_j^{(t)}$, the pretrained detector predicts a set of 3D bounding boxes:

$$\hat{\mathcal{B}}_j^{(t)} = \psi_{\mathrm{DET}}\left(y_j^{(t)}\right), \qquad \mathcal{B}_j^{\mathrm{gt}(t)} \text{ denotes the corresponding ground-truth set.}$$

Per-frame detection metrics (mAP and NDS) are computed following [58, 66] using standard matching and error terms. The dataset-level 3D detection score averages these values across all generated frames:

$$\boxed{\mathcal{S}_{\mathrm{Det}}(\mathcal{Y}) = \tfrac{1}{N_g T} \sum_{j=1}^{N_g} \sum_{t=1}^{T} \mathrm{NDS}\left(y_j^{(t)}\right)} \tag{20}$$

Higher $\mathcal{S}_{\mathrm{Det}}$ (and mAP) indicates that the generated frames support more accurate 3D reasoning and reliable downstream perception for autonomous driving.

10.2.3 Implementation Details

The 3D detection evaluation uses the same pretrained BEVFusion model as in Section 10.1, with its detection head producing 3D bounding boxes on the nuScenes BEV range $[-51.2, 51.2]$ m^2 and a 150 × 150 grid. Predicted boxes are evaluated against ground-truth annotations using standard nuScenes 3D detection metrics.

10.2.4 Examples

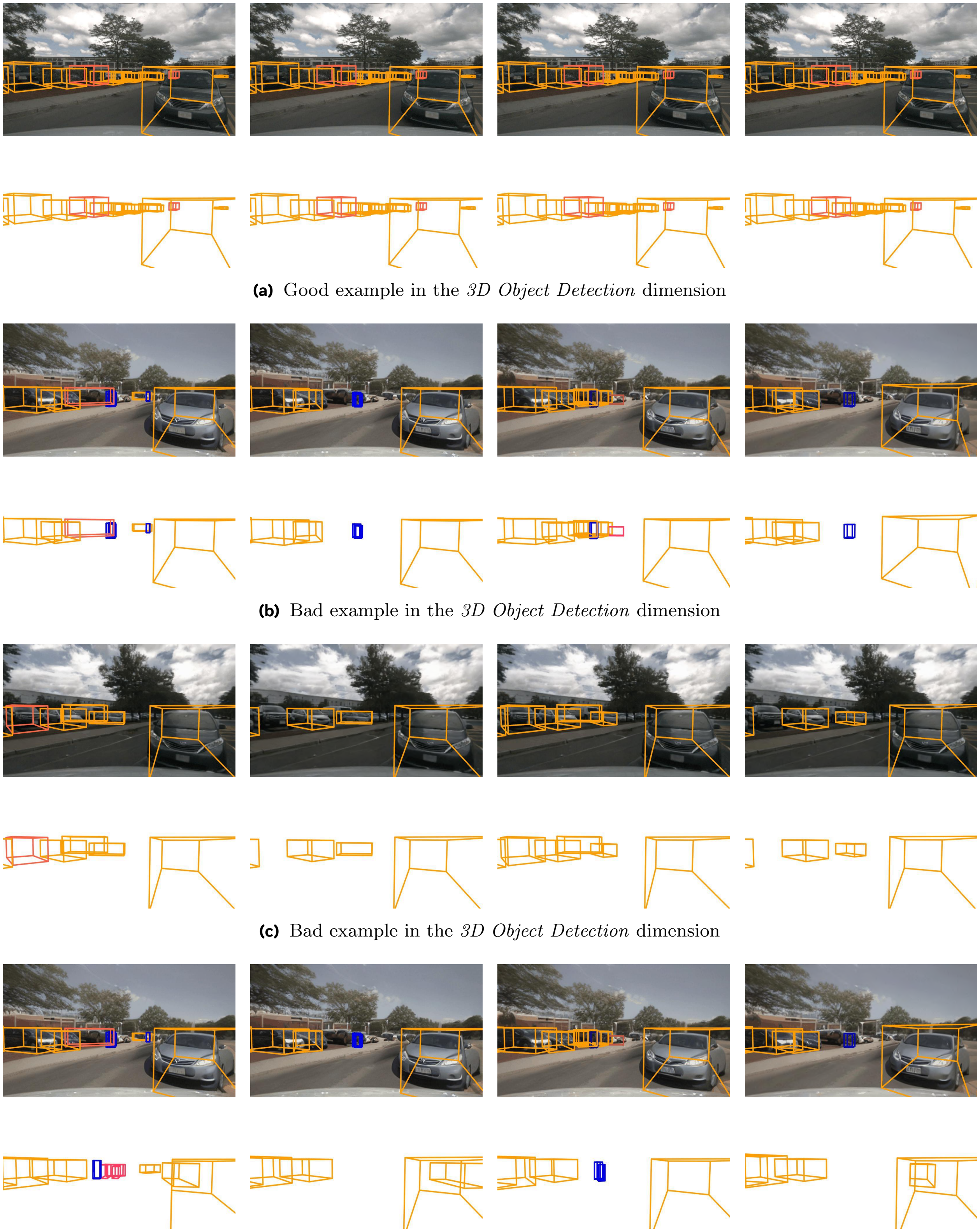
Figure 26 provides typical examples of videos with good and bad quality in terms of *3D Object Detection*.

10.2.5 Evaluation & Analysis

Table 21 provides the complete results of models in terms of *3D Object Detection*.

Table 21 Complete comparisons of state-of-the-art driving world models in terms of *3D Object Detection* in WorldLens.

$\mathcal{S}_{\mathrm{Det}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
mAP (↑)	0.1178	0.1636	0.1961	0.944	0.2242	0.1562	0.3657
mATE (↓)	0.9435	0.9469	0.8443	1.0354	0.9256	0.8870	0.7356
mASE (↓)	0.3400	0.3207	0.3273	0.3479	0.3214	0.3218	0.2919
mAOE (↓)	0.7834	0.8237	0.5930	0.7734	0.5252	0.6509	0.4400
mAVE (↓)	1.0133	0.8039	0.8904	0.8629	0.7897	0.7061	0.6821
mAAE (↓)	0.2814	0.2520	0.2349	0.2917	0.2374	0.2265	0.2072
NDS (↑)	0.2241	0.2671	0.3090	0.2196	0.3322	0.2989	0.4472



(a) Good example in the *3D Object Detection* dimension

(b) Bad example in the *3D Object Detection* dimension

(c) Bad example in the *3D Object Detection* dimension

(d) Bad example in the *3D Object Detection* dimension

Figure 26 Examples of "good" and "bad" downstream task performances in terms of *3D Object Detection* in WorldLens.

10.3 3D Object Tracking

10.3.1 Definition

3D Object Tracking evaluates whether generated videos preserve consistent object motion and identity information that supports temporal data association. A pretrained tracker $\psi_{\text{TRK}}(\cdot)$, trained on ground-truth sequences, is applied to each generated video to estimate 3D trajectories of dynamic objects.

Following the nuScenes tracking protocol [9], tracking performance is measured using the Average Multi-Object Tracking Accuracy (AMOTA), which integrates precision, recall, and association quality across recall thresholds. Higher AMOTA values indicate that the generated videos exhibit realistic temporal dynamics, enabling stable object tracking over time.

10.3.2 Formulation

For each generated video $y_j = \{y_j^{(t)}\}_{t=1}^{T}$, the tracker predicts a set of object trajectories:

$$\hat{\mathcal{T}}_j = \psi_{\text{TRK}}(y_j) = \{\hat{\tau}_n = \{\hat{\mathbf{b}}_n^{(t)}\}_{t \in \mathcal{I}_n}\}_{n=1}^{N_{\text{trk}}},$$

and $\mathcal{T}_j^{\text{gt}}$ denotes the corresponding ground-truth trajectories. Tracking accuracy is evaluated using the official nuScenes metrics [9], including MOTA and AMOTA, where higher scores indicate more reliable data association and motion continuity.

The dataset-level 3D tracking metric averages per-video AMOTA over all generated sequences:

$$\boxed{\mathcal{S}_{\text{Trk}}(\mathcal{Y}) = \tfrac{1}{N_g} \sum_{j=1}^{N_g} \text{AMOTA}(y_j)} \qquad (21)$$

Higher $\mathcal{S}_{\text{Trk}}$ indicates that generated videos maintain realistic and temporally coherent object motion, supporting accurate long-term tracking.

10.3.3 Implementation Details

We evaluate the 3D object tracking performance using the pretrained camera-only ADA-Track [22], following its official nuScenes configuration. The tracker is run directly on the generated multi-view videos.

10.3.4 Examples

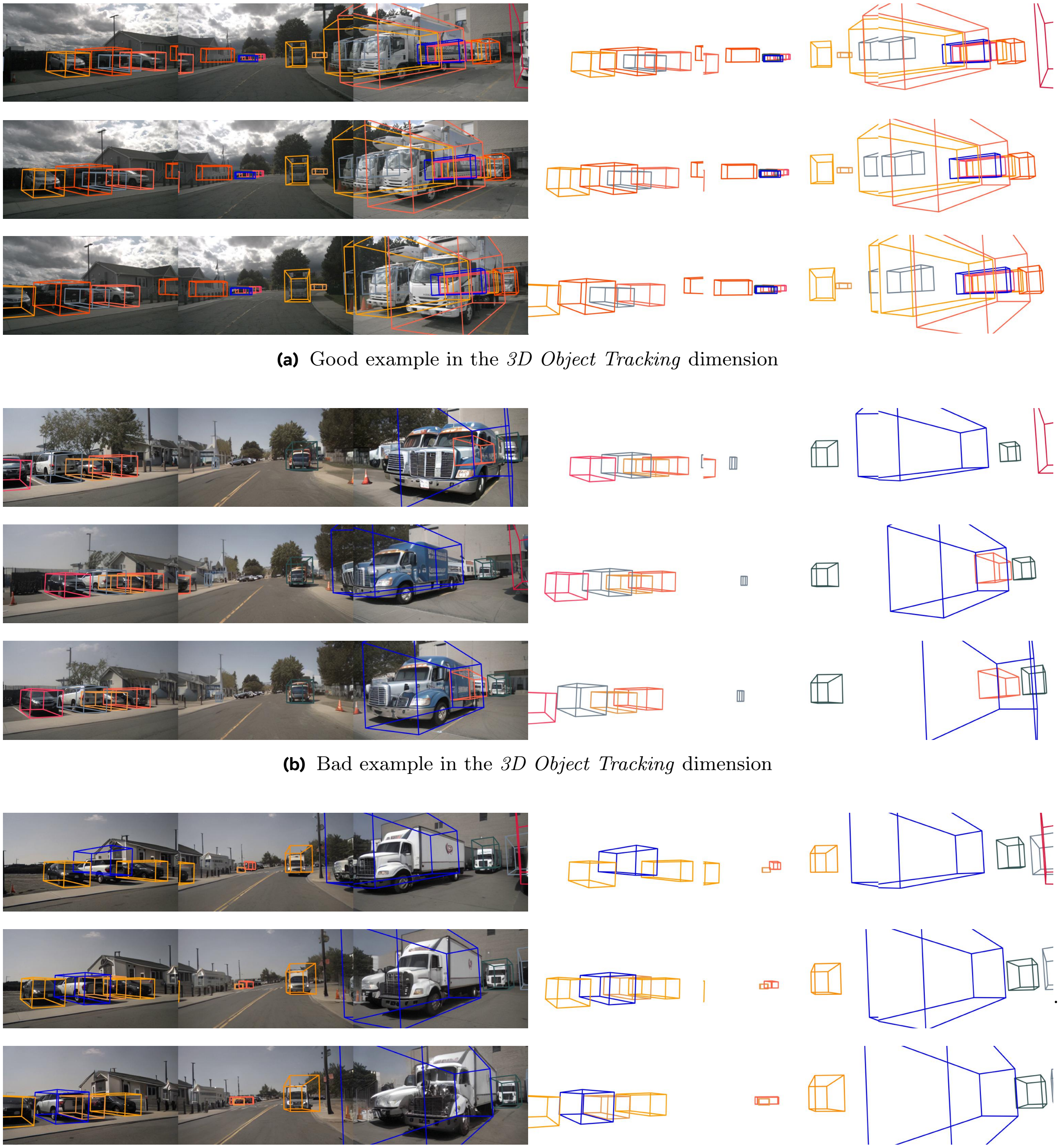
Figure 27 provides typical examples of videos with good and bad quality in terms of *3D Object Tracking*.

10.3.5 Evaluation & Analysis

Table 22 provides the complete results of models in terms of *3D Object Tracking*.

Table 22 Complete comparisons of state-of-the-art driving world models in terms of *3D Object Tracking* in WorldLens.

$\mathcal{S}_{\text{Trk}}(\cdot)$	**MagicDrive**	**DreamForge**	**DriveDreamer-2**	**OpenDWM**	**DiST-4D**	**$\mathcal{X}$-Scene**	**Empirical**
	[ICLR'24]	[arXiv'24]	[AAAI'25]	[CVPR'25]	[ICCV'25]	[NeurIPS'25]	**Max**
AMOTP (↓)	1.843	1.812	1.778	1.848	1.705	1.829	1.405
Recall (↑)	15.50%	16.10%	19.50%	12.40%	29.20%	15.56%	45.30%
MOTA (↑)	8.70%	9.70%	12.70%	7.40%	15.10%	10.60%	34.30%
FN (↓)	5093	4622	4361	4820	3725	4799	2678
TP (↑)	1615	2084	2343	1887	2977	1909	4027
AMOTA (↑)	7.90%	10.30%	13.30%	6.90%	15.30%	8.80%	36.30%



(a) Good example in the *3D Object Tracking* dimension

(b) Bad example in the *3D Object Tracking* dimension

(c) Bad example in the *3D Object Tracking* dimension

Figure 27 Examples of "good" and "bad" downstream task performances in terms of *3D Object Tracking* in WorldLens.

10.4 Occupancy Prediction

10.4.1 Definition

Occupancy Prediction evaluates whether generated videos enable accurate 3D reconstruction of scene geometry and semantics. We adopt the RayIoU metric [96], which measures semantic and geometric agreement *along camera rays* rather than voxel overlap. For each ray, RayIoU compares the *frontmost* occupied voxel in the predicted and ground-truth volumes, requiring both class correctness and depth proximity within a tolerance δ. This ray-wise formulation avoids the depth-ambiguity of voxel mIoU (which may reward thick surfaces) and naturally supports multi-pose scene completion evaluation via ray casting.

10.4.2 Formulation

A frozen occupancy estimator $\psi_{\text{Occ}}(\cdot)$ predicts a probabilistic 3D volume for each generated video:

$$\hat{\mathbf{O}}_j = \psi_{\text{Occ}}(y_j)\ , \qquad \hat{\mathbf{O}}_j \in [0,1]^{X \times Y \times Z}\ .$$

Let $\mathcal{R}$ denote the set of sampled query rays (with distance-balanced resampling). For each ray $r \in \mathcal{R}$, denote the frontmost occupied voxel in prediction and ground truth by $(\hat{d}_r, \hat{c}_r)$ and $(d_r^{\text{gt}}, c_r^{\text{gt}})$. A prediction is correct if $\hat{c}_r = c_r^{\text{gt}}$ and $|\,\hat{d}_r - d_r^{\text{gt}}\,| \le \delta$. The RayIoU at tolerance δ is defined as:

$$\text{RayIoU@}\delta = \tfrac{1}{C} \sum\nolimits_{c=1}^{C} \tfrac{\text{TP}_c(\delta)}{\text{TP}_c(\delta)+\text{FP}_c(\delta)+\text{FN}_c(\delta)}\ ,$$

and the mean RayIoU (mRayIoU) aggregates multiple tolerances:

$$\text{mRayIoU} = \tfrac{1}{3} \sum\nolimits_{\delta \in \{1,2,4\}} \text{RayIoU@}\delta\ .$$

The dataset-level semantic occupancy score averages mRayIoU across all generated videos:

$$\boxed{\mathcal{S}_{\text{Occ}}(\mathcal{Y}) = \tfrac{1}{N_g} \sum\nolimits_{j=1}^{N_g} \text{mRayIoU}(y_j)} \tag{22}$$

Higher $\mathcal{S}_{\text{Occ}}$ indicates that generated scenes enable more accurate, depth-consistent, and semantically faithful occupancy reconstruction.

10.4.3 Implementation Details

We perform occupancy prediction using the pretrained SparseOcc model [96]. The model is applied to the generated multi-view frames following the official nuScenes camera-only configuration, producing voxel-wise semantic occupancy grids within a $[-40, 40] \times [-40, 40] \times [-1, 5.4]$ m 3D volume for evaluation.

10.4.4 Examples

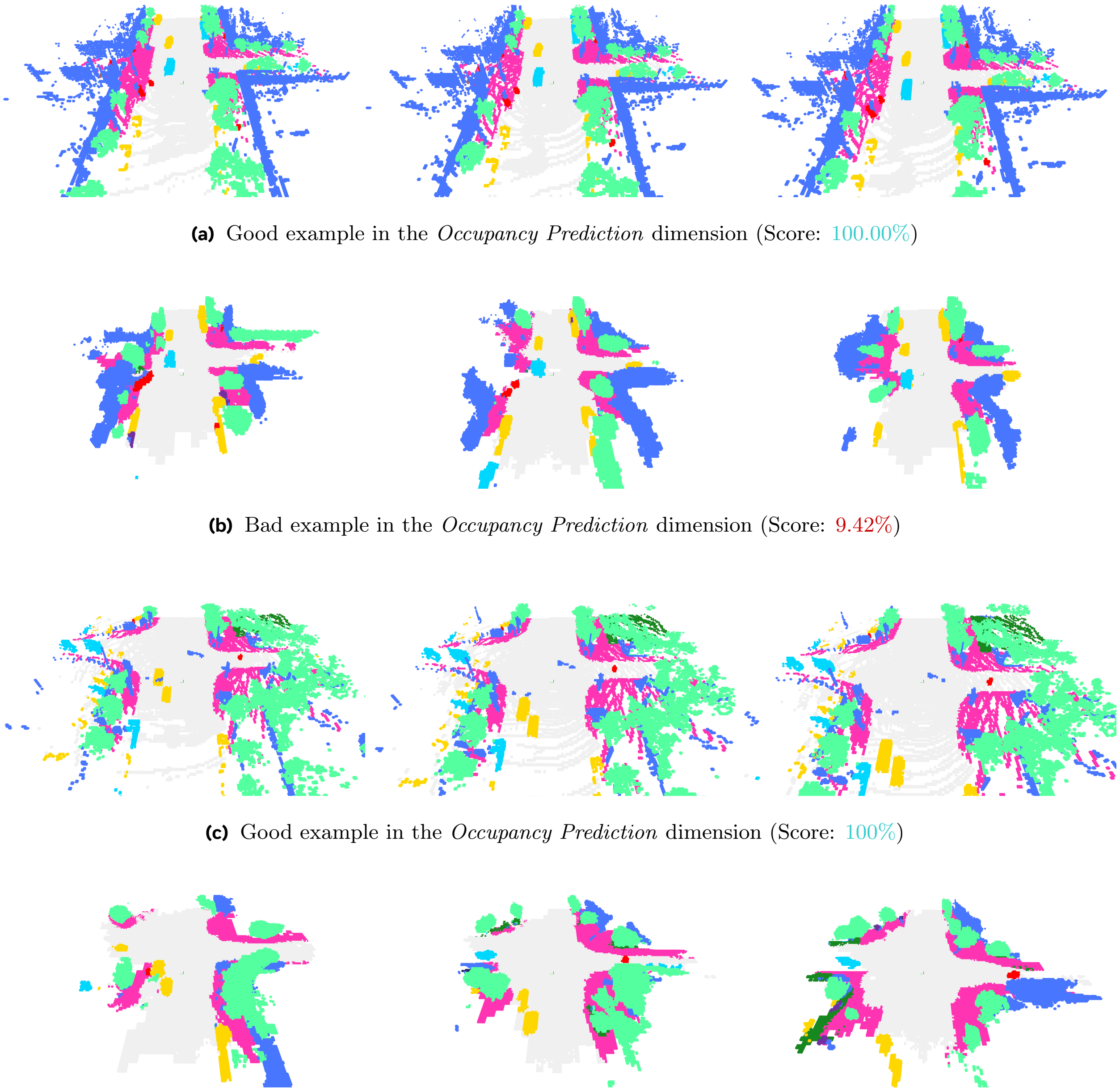
Figure 28 provides typical examples of videos with good and bad quality in terms of *Occupancy Prediction*.

10.4.5 Evaluation & Analysis

Table 23 provides the complete results of models in terms of *Occupancy Prediction*.

Table 23 Complete comparisons of state-of-the-art driving world models in terms of *Occupancy Prediction* in WorldLens.

$\mathcal{S}_{\text{Occ}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
RayIoU@1m (↑)	17.24%	18.24%	20.32%	17.78%	18.81%	18.00%	29.04%
RayIoU@2m (↑)	23.53%	24.10%	27.36%	25.35%	26.50%	23.93%	38.17%
RayIoU@4m (↑)	28.65%	28.79%	32.77%	31.32%	33.00%	29.12%	43.93%
Average (↑)	23.14%	23.71%	26.82%	24.82%	26.10%	23.68%	37.05%



(a) Good example in the *Occupancy Prediction* dimension (Score: 100.00%)

(b) Bad example in the *Occupancy Prediction* dimension (Score: 9.42%)

(c) Good example in the *Occupancy Prediction* dimension (Score: 100%)

(d) Bad example in the *Occupancy Prediction* dimension (Score: 11.27%)

Figure 28 Examples of “good” and “bad” downstream task performances in terms of *Occupancy Prediction* in WorldLens.

11 Aspect 5: Human Preference

This section presents human-centered evaluations. While quantitative measures capture specific aspects of fidelity, consistency, and geometric accuracy, they cannot fully reflect **how humans perceive realism, stability, and overall scene quality**. To bridge this gap, we introduce a human preference study that scores generated videos across multiple dimensions, providing a holistic and perceptually grounded assessment of model performance.

11.1 World Realism – Overall Realism

Overall Realism measures the global visual believability. Annotators judge whether the generated video "looks like a real-world driving recording". They are instructed to judge each clip based on the following criteria:

- Structural and perspective coherence of the environment.
- Visual stability without severe flicker, tearing, or geometric warping.
- Realistic lighting, shadows, and surface textures.
- Consistent composition of static (roads, buildings, sky) and dynamic (vehicles, pedestrians) elements.

Higher *Overall Realism* indicates that generated scenes are globally coherent, visually stable, and perceptually indistinguishable from real-world videos.

11.1.1 Protocol

Each generated video is rated on a 1–10 scale of perceived realism:

Score	Level	Description
1	Extremely Unrealistic	Severe structural and temporal defects; global flicker or collapse of scene; lighting/texture incoherent; scene immediately identifiable as synthetic.
3	Unrealistic	Local artifacts such as inconsistent textures, ghosting, or unstable motion, but the overall layout remains interpretable.
5	Moderately Realistic	Global appearance mostly coherent with minor motion discontinuities or soft blur; realism is partially convincing.
7	Realistic	Scene composition, motion continuity, and lighting are natural; just some small imperfections but do not affect perceived realism.
9	Highly Realistic	Scene fully photorealistic in both space and time; almost indistinguishable from real-world footage by human eyes.
10	Ground Truth	-

11.1.2 Examples

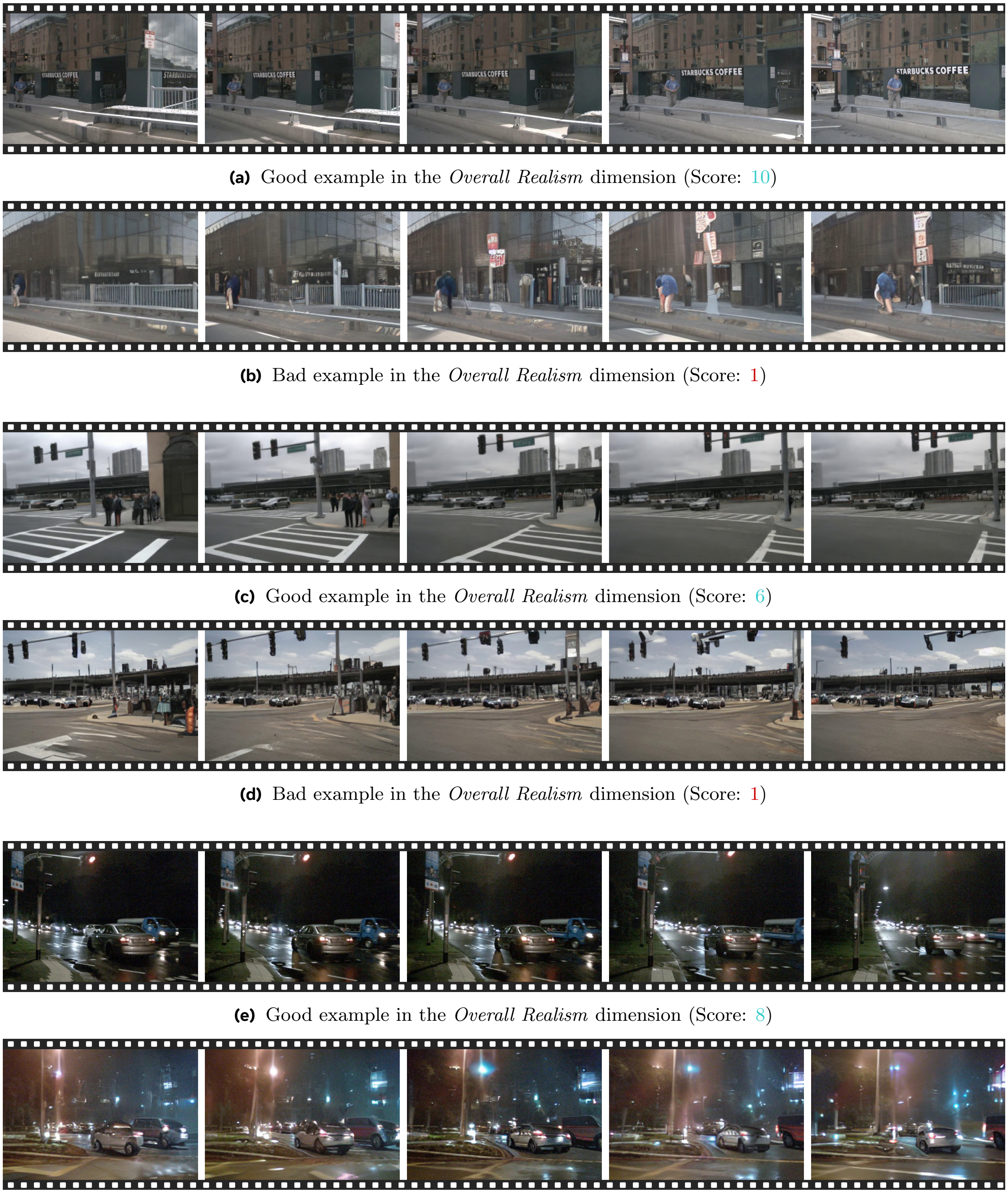
Figure 29 provides typical examples of videos with good and bad quality in terms of *Overall Realism.*

11.1.3 Evaluation & Analysis

Table 24 provides the complete results of models in terms of *Overall Realism.*

Table 24 Complete comparisons of state-of-the-art driving world models in terms of *Overall Realism* in WorldLens.

Overall Realism	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
min	2.0	2.0	2.0	2.0	2.0	2.0	-
max	4.0	6.0	6.0	6.0	6.0	4.0	-
mean	2.062	2.204	2.256	2.209	2.320	2.080	10
std	0.347	0.620	0.865	0.801	0.912	0.392	-
median	2.0	2.0	2.0	2.0	2.0	2.0	-
q25	2.0	2.0	2.0	2.0	2.0	2.0	-
q75	2.0	2.0	2.0	2.0	2.0	2.0	-



(a) Good example in the *Overall Realism* dimension (Score: 10)

(b) Bad example in the *Overall Realism* dimension (Score: 1)

(c) Good example in the *Overall Realism* dimension (Score: 6)

(d) Bad example in the *Overall Realism* dimension (Score: 1)

(e) Good example in the *Overall Realism* dimension (Score: 8)

(f) Bad example in the *Overall Realism* dimension (Score: 1)

Figure 29 Examples of "good" and "bad" human preference alignments in terms of *Overall Realism* in WorldLens.

11.2 World Realism - Vehicle Realism

Vehicle Realism isolates the perceptual authenticity of vehicles within the scene, focusing solely on their visual appearance. Annotators evaluate whether vehicles "look like real cars". Annotators are instructed to judge each clip according to the following criteria:

- Correct body shape, door/roof/wheel-arch proportions, and stable contours without deformation.
- Realistic metallic paint, plastic, glass, tires, and recognizable small components (logos, grilles, lamps).
- Natural highlights, shadows, and reflections under various weather and illumination conditions.
- Color, texture, and boundary stability across adjacent frames.

High *Vehicle Realism* reflects consistent car geometry, convincing materials, physically plausible reflections, and temporally stable rendering. Low scores correspond to warped, "rubber-like" cars with flickering colors, melted textures, or incoherent lighting.

11.2.1 Protocol

Each generated video is rated on a 1–10 scale of perceived realism:

Score	**Level**	**Description**
1	Extremely Unrealistic	Vehicle geometry or texture severely distorted; missing parts, collapsed meshes, or flickering silhouettes; clearly fake appearance.
3	Unrealistic	Vehicles roughly shaped but show color inconsistency, unstable reflections, or unnatural motion patterns.
5	Moderately Realistic	Vehicles recognizable with mostly correct proportions and materials; small surface or temporal noise visible.
7	Realistic	Vehicle shape, motion, and illumination coherent and stable; only have some minor local imperfections.
9	Highly Realistic	Fully natural vehicles with correct proportions, lighting response, and dynamic reflections; seamlessly integrated in the scene.
10	Ground Truth	-

11.2.2 Examples

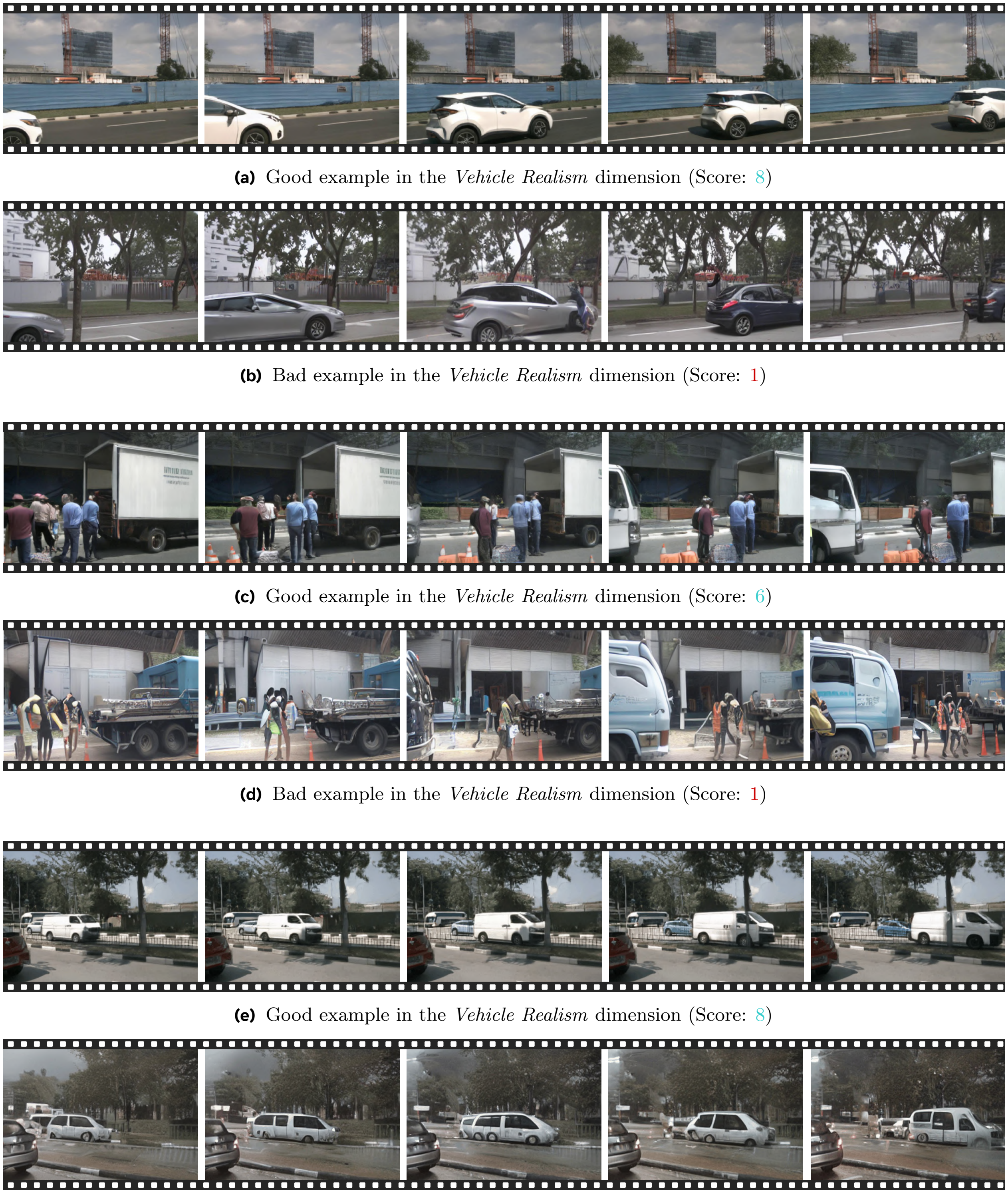
Figure 30 provides typical examples of videos with good and bad quality in terms of *Vehicle Realism*.

11.2.3 Evaluation & Analysis

Table 25 provides the complete results of models in terms of *Vehicle Realism*.

Table 25 Complete comparisons of state-of-the-art driving world models in terms of *Vehicle Realism* in WorldLens.

Vehicle Realism	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
min	2.0	2.0	2.0	2.0	2.0	2.0	-
max	4.0	6.0	8.0	8.0	8.0	8.0	-
mean	2.036	2.043	2.720	2.757	2.328	2.216	10
std	0.268	0.328	1.700	1.584	1.011	0.808	-
median	2.0	2.0	2.0	2.0	2.0	2.0	-
q25	2.0	2.0	2.0	2.0	2.0	2.0	-
q75	2.0	2.0	2.0	2.0	2.0	2.0	-



(a) Good example in the *Vehicle Realism* dimension (Score: 8)

(b) Bad example in the *Vehicle Realism* dimension (Score: 1)

(c) Good example in the *Vehicle Realism* dimension (Score: 6)

(d) Bad example in the *Vehicle Realism* dimension (Score: 1)

(e) Good example in the *Vehicle Realism* dimension (Score: 8)

(f) Bad example in the *Vehicle Realism* dimension (Score: 1)

Figure 30 Examples of "good" and "bad" human preference alignments in terms of *Vehicle Realism* in WorldLens.

11.3 World Realism - Pedestrian Realism

Pedestrian Realism measures whether humans in generated videos look and move like real people. It focuses on anatomical plausibility, natural appearance, and temporal stability of pedestrians. Annotators are instructed to judge each clip according to the following criteria:

- Realistic head-torso-limb ratios, joint positions, and poses without twisted or intersecting limbs.
- Plausible garment structure, texture clarity, and consistency of accessories.
- Smooth and natural shading without wax-like or distorted faces.
- Continuous appearance without flickering, sliding, or sudden disappearance.
- Whether pedestrians resemble real filmed humans rather than avatars or composites.

Higher *Pedestrian Realism* indicates pedestrians with stable body structures, coherent motion, realistic textures, and natural temporal behavior.

11.3.1 Protocol

Each generated video is rated on a 1–10 scale of perceived realism:

Score	Level	Description
1	Extremely Unrealistic	Human figures deformed or incomplete; limbs twisted or merged; motion violates body mechanics; instantly recognizable as artificial.
3	Unrealistic	Human silhouettes intact but with visible motion or shape glitches, coarse skin/cloth texture, or flicker; gait unnatural.
5	Moderately Realistic	Pedestrians generally human-like with slight stiffness or occasional temporal instability, shape distortions, or texture issues.
7	Realistic	Natural body proportions, coherent motion, and stable clothing appearance; plausible human dynamics.
9	Highly Realistic	Anatomically and kinematically accurate humans with smooth gait, fine-grained details, and temporally consistent appearance.
10	Ground Truth	-

11.3.2 Examples

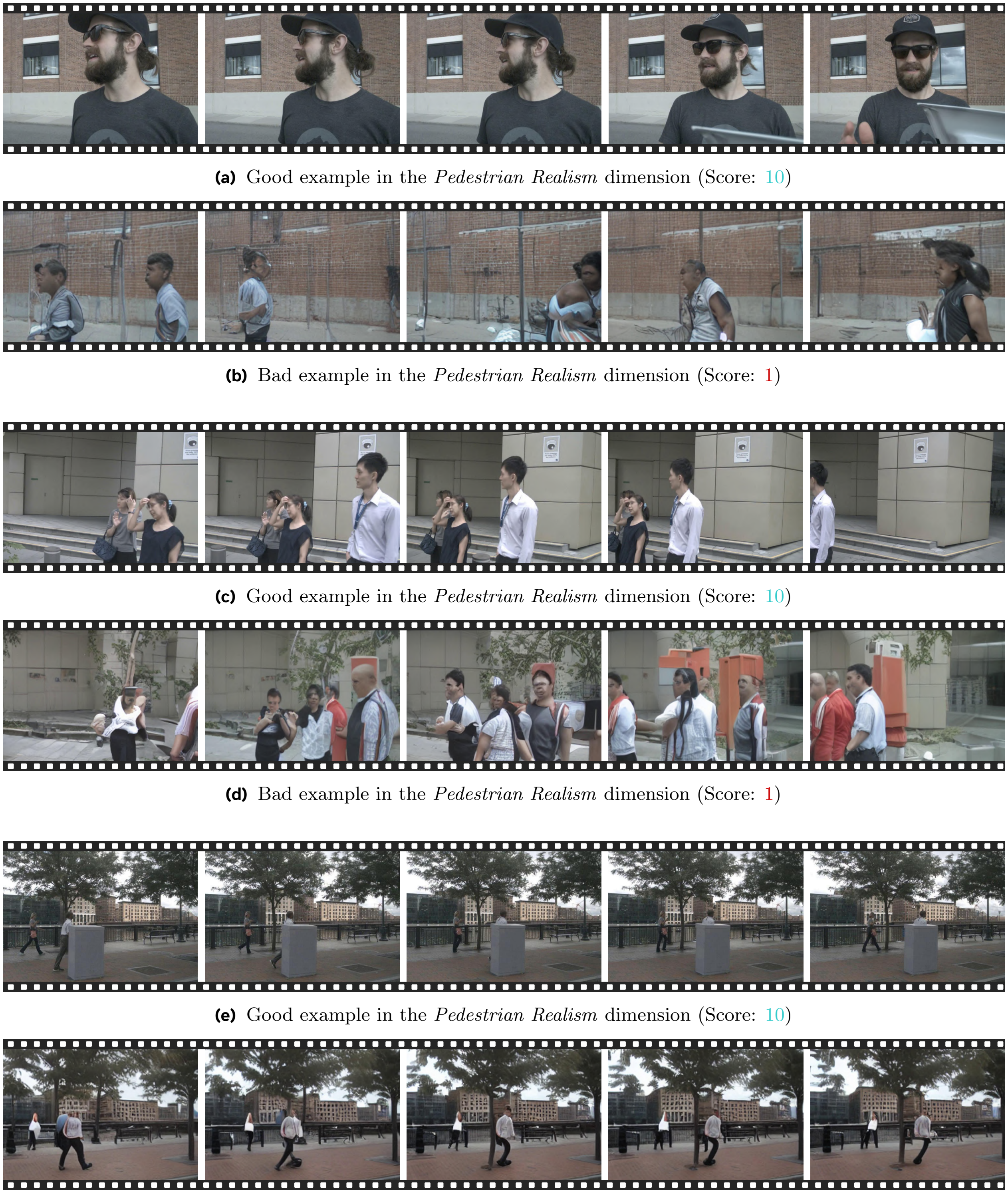
Figure 31 provides typical examples of videos with good and bad quality in terms of *Pedestrian Realism.*

11.3.3 Evaluation & Analysis

Table 26 provides the complete results of models in terms of *Pedestrian Realism.*

Table 26 Complete comparisons of state-of-the-art driving world models in terms of *Pedestrian Realism* in WorldLens.

Pedestrian Realism	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
min	2.0	2.0	2.0	2.0	2.0	2.0	-
max	4.0	6.0	6.0	4.0	6.0	4.0	-
mean	2.288	2.352	2.341	2.325	2.406	2.293	10
std	0.618	0.727	0.703	0.671	0.832	0.629	-
median	2.0	2.0	2.0	2.0	2.0	2.0	-
q25	2.0	2.0	2.0	2.0	2.0	2.0	-
q75	2.0	2.0	2.0	2.0	2.0	2.0	-



(a) Good example in the *Pedestrian Realism* dimension (Score: 10)

(b) Bad example in the *Pedestrian Realism* dimension (Score: 1)

(c) Good example in the *Pedestrian Realism* dimension (Score: 10)

(d) Bad example in the *Pedestrian Realism* dimension (Score: 1)

(e) Good example in the *Pedestrian Realism* dimension (Score: 10)

(f) Bad example in the *Pedestrian Realism* dimension (Score: 1)

Figure 31 Examples of "good" and "bad" human preference alignments in terms of *Pedestrian Realism* in WorldLens.

11.4 Physical Plausibility

Physical Plausibility evaluates whether the motions, interactions, and visual evolution of a generated driving scene are consistent with basic physical laws and causal structure in the real world. This dimension explicitly targets *physics and dynamics*: whether objects move, collide, occlude, and respond in ways that respect continuity, inertia, contact, and depth ordering. Annotators are instructed to judge each clip according to the following criteria:

- Positions, velocities, colors, and textures should evolve smoothly over time, without teleportation, duplication, spontaneous appearance or disappearance, or violent jumps in shape or brightness.
- Vehicles, pedestrians, and static elements (barriers, poles, buildings) should not interpenetrate. Feet should visually remain on the ground when walking, and objects should not float or sink into surfaces.
- Foreground and background elements should obey consistent occlusion relationships. Distant objects should not suddenly occlude closer ones, and elements like traffic lights, fences, and signboards should not phase through other geometry.
- Highlights, reflections, glare, and shadows should change smoothly with camera motion and object movement, without unexplained flashes, patches of incoherent reflection, or abrupt brightness jumps.

Higher *Physical Plausibility* indicates that generated worlds exhibit more physically consistent dynamics.

11.4.1 Protocol

Each generated video is rated on a 1–10 scale of perceived realism:

Score	Level	Description
1	Extremely Implausible	Frequent physics violations: teleportation, penetration, inconsistent occlusion, abrupt geometry or lighting jumps.
3	Implausible	Localized but noticeable non-physical events (*e.g.*, a single penetration or transient reflection anomaly); scene remains somewhat coherent.
5	Moderately Plausible	Motion and contact mostly realistic with occasional small violations (*e.g.*, light flicker, minor occlusion inversion); perceptually acceptable but imperfect.
7	Plausible	Motion, occlusion, and lighting largely follow physical laws; minor irregularities remain in non-critical regions.
9	Highly Plausible	Entire clip adheres to continuity, contact, reflection, and causality constraints; fully consistent with real-world physics.
10	Ground Truth	-

11.4.2 Examples

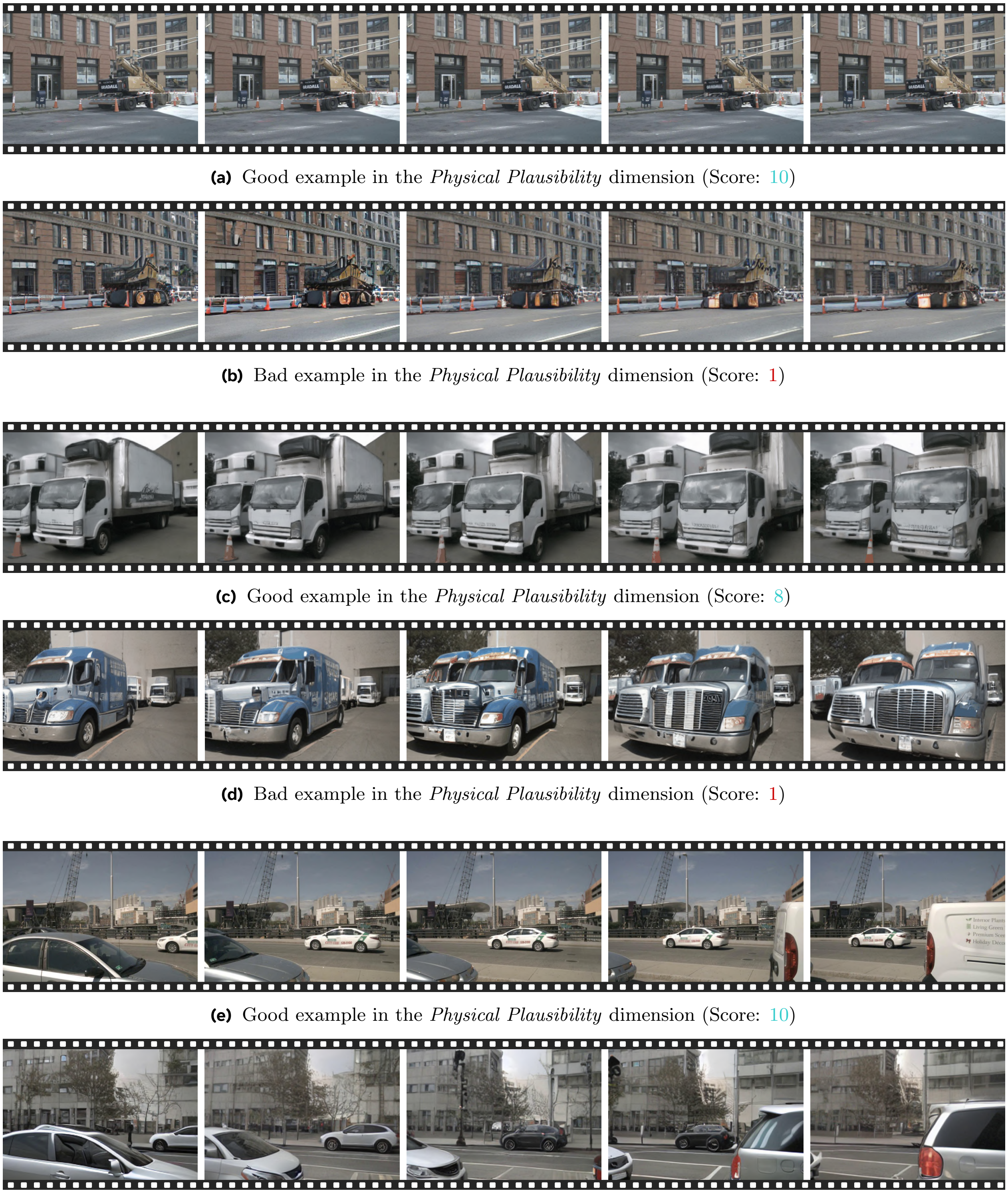
Figure 32 provides typical examples of videos with good and bad quality in terms of *Physical Plausibility*.

11.4.3 Evaluation & Analysis

Table 27 provides the complete results of models in terms of *Physical Plausibility*.

Table 27 Complete comparisons of state-of-the-art driving world models in terms of *Physical Plausibility* in WorldLens.

Physical Plausibility	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
min	2.0	2.0	2.0	2.0	2.0	2.0	-
max	4.0	4.0	8.0	8.0	10.0	4.0	-
mean	2.300	2.300	2.380	2.312	2.583	2.292	10
std	0.640	0.640	0.783	0.674	1.187	0.626	-
median	2.0	2.0	2.0	2.0	2.0	2.0	-
q25	2.0	2.0	2.0	2.0	2.0	2.0	-
q75	2.0	2.0	2.0	2.0	3.0	2.0	-



(a) Good example in the *Physical Plausibility* dimension (Score: 10)

(b) Bad example in the *Physical Plausibility* dimension (Score: 1)

(c) Good example in the *Physical Plausibility* dimension (Score: 8)

(d) Bad example in the *Physical Plausibility* dimension (Score: 1)

(e) Good example in the *Physical Plausibility* dimension (Score: 10)

(f) Bad example in the *Physical Plausibility* dimension (Score: 1)

Figure 32 Examples of "good" and "bad" human preference alignments in terms of *Physical Plausibility* in WorldLens.

11.5 3D & 4D Consistency

Physical Plausibility measures how well the 3D structure and temporal evolution of objects in a generated video align with those in the corresponding real (ground-truth) sequence. Rather than judging raw pixels, this dimension focuses on the stability and accuracy of 3D bounding boxes over time, as estimated by a pretrained tracking or detection model applied to both generated and real videos. Annotators are instructed to judge each clip according to the following criteria:

- For each object, the 3D box size, orientation, and position should evolve smoothly over time, without jitter, sudden jumps, unnatural scaling, or misalignment with the underlying object.
- Tracks should persist as long as the object is visible, without frequent flickering, disappearing-and-reappearing, or drifting away from the target.
- The number and spatial arrangement of boxes in the generated view should broadly match those in the ground-truth view, especially for prominent nearby objects.
- In dynamic scenes, the motion direction and speed of boxes in the generated view should be close to those in the ground-truth view; in static scenes, boxes should remain essentially still.

Higher scores indicate that 3D box trajectories in generated videos closely track those in real scenes.

11.5.1 Protocol

Each generated video is rated on a 1–10 scale of perceived realism:

Score	Level	Description
1	Extremely Inconsistent	3D boxes unstable or mismatched; severe jitter, missing frames, or large misalignment from the ground truth.
3	Inconsistent	Rough trajectory trend visible but with clear instability or mismatched counts; large misalignment from the ground truth.
5	Moderately Consistent	3D boxes mostly aligned with ground truth; however, there are still minor jitter or missing boxes that can be detected by human eyes.
7	Consistent	Smooth, stable trajectories and coherent spatial alignment; only slight temporal noise; the number of detected 3D boxes mostly aligns with the ground truth.
9	Highly Consistent	Generated 3D boxes match ground-truth positions and motions almost perfectly across time; the inconsistency can hardly be detected by human eyes.
10	Ground Truth	-

11.5.2 Examples

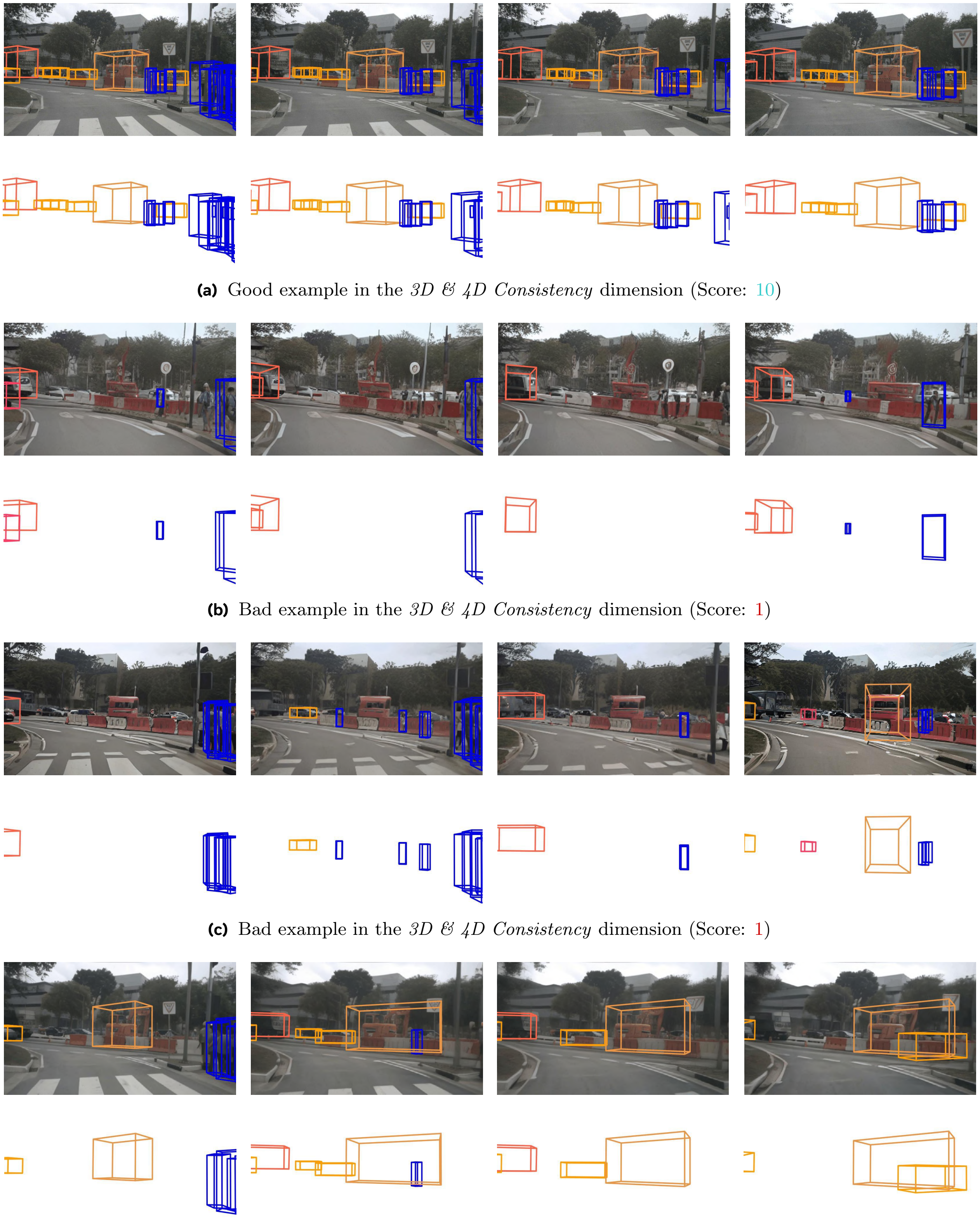
Figure 33 provides typical examples of videos with good and bad quality in terms of *3D & 4D Consistency.*

11.5.3 Evaluation & Analysis

Table 28 provides the complete results of models in terms of *3D & 4D Consistency.*

Table 28 Complete comparisons of state-of-the-art driving world models in terms of *3D & 4D Consistency* in WorldLens.

3D & 4D Consistency	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
min	2.0	2.0	2.0	2.0	2.0	2.0	-
max	10.0	10.0	10.0	10.0	10.0	10.0	-
mean	2.455	2.743	2.751	2.920	2.961	2.431	10
std	1.061	1.378	1.530	1.467	1.405	1.161	-
median	2.0	2.0	2.0	2.0	2.0	2.0	-
q25	2.0	2.0	2.0	2.0	2.0	2.0	-
q75	2.0	4.0	3.0	4.0	4.0	2.0	-



(a) Good example in the *3D & 4D Consistency* dimension (Score: 10)

(b) Bad example in the *3D & 4D Consistency* dimension (Score: 1)

(c) Bad example in the *3D & 4D Consistency* dimension (Score: 1)

(d) Bad example in the *3D & 4D Consistency* dimension (Score: 1)

Figure 33 Examples of "good" and "bad" human preference alignments in terms of *3D & 4D Consistency* in WorldLens.

11.6 Behavioral Safety

Behavioral Safety measures how safe and predictable the visible behavior of traffic participants appears in generated driving videos, as judged by human observers. Rather than evaluating visual realism alone, this dimension focuses on whether vehicles, pedestrians, cyclists, and other agents interact with each other and with key scene elements in a way that is consistent with basic traffic rules and low-risk driving. Annotators are instructed to judge each clip according to the following criteria:

- Obvious impossible behaviors, *e.g.*, sudden teleportation, splitting or merging of agents, agents appearing or disappearing without cause, or severe shape deformation that destroys basic spatial relations.
- Whether agent behavior clearly contradicts prominent traffic signals, signs, or lane markings (for example, ignoring a red light, driving against traffic, or violating stop or yield indications).
- Whether vehicles and road users maintain reasonable gaps, avoid implausible near-collisions or illegal crossings, and follow trajectories that are smooth and predictable rather than erratic or conflict-prone.
- Whether scene distortions, flickering, or object deformations directly impair the ability to read safety-critical cues, such as lane boundaries, signal states, and relative positions of agents.

Higher *Behavioral Safety* indicates that generated videos tend to display traffic behavior that raters judge as safe and consistent with basic road rules.

11.6.1 Protocol

Each generated video is rated on a 1–10 scale of perceived realism:

Score	Level	Description
1	Extremely Unsafe	Catastrophic anomalies or impossible events (teleportation, splitting, collisions, severe signal violations) that make the scene unsafe at a glance.
3	Mostly Unsafe	Partial or localized safety violations, *e.g.*, brief teleportation, collisions, or unreadable signal states, that clearly degrade perceived safety.
5	Moderately Safe	Generally safe behavior with mild instability or motion artifacts; no major conflicts, though realism is limited.
7	Safe	Predictable and stable behavior; vehicles and pedestrians maintain proper spacing and respect traffic logic.
9	Highly Safe	Fully natural and rule-abiding behavior; smooth trajectories, compliant with signals, and entirely risk-free appearance.
10	Ground Truth	-

11.6.2 Examples

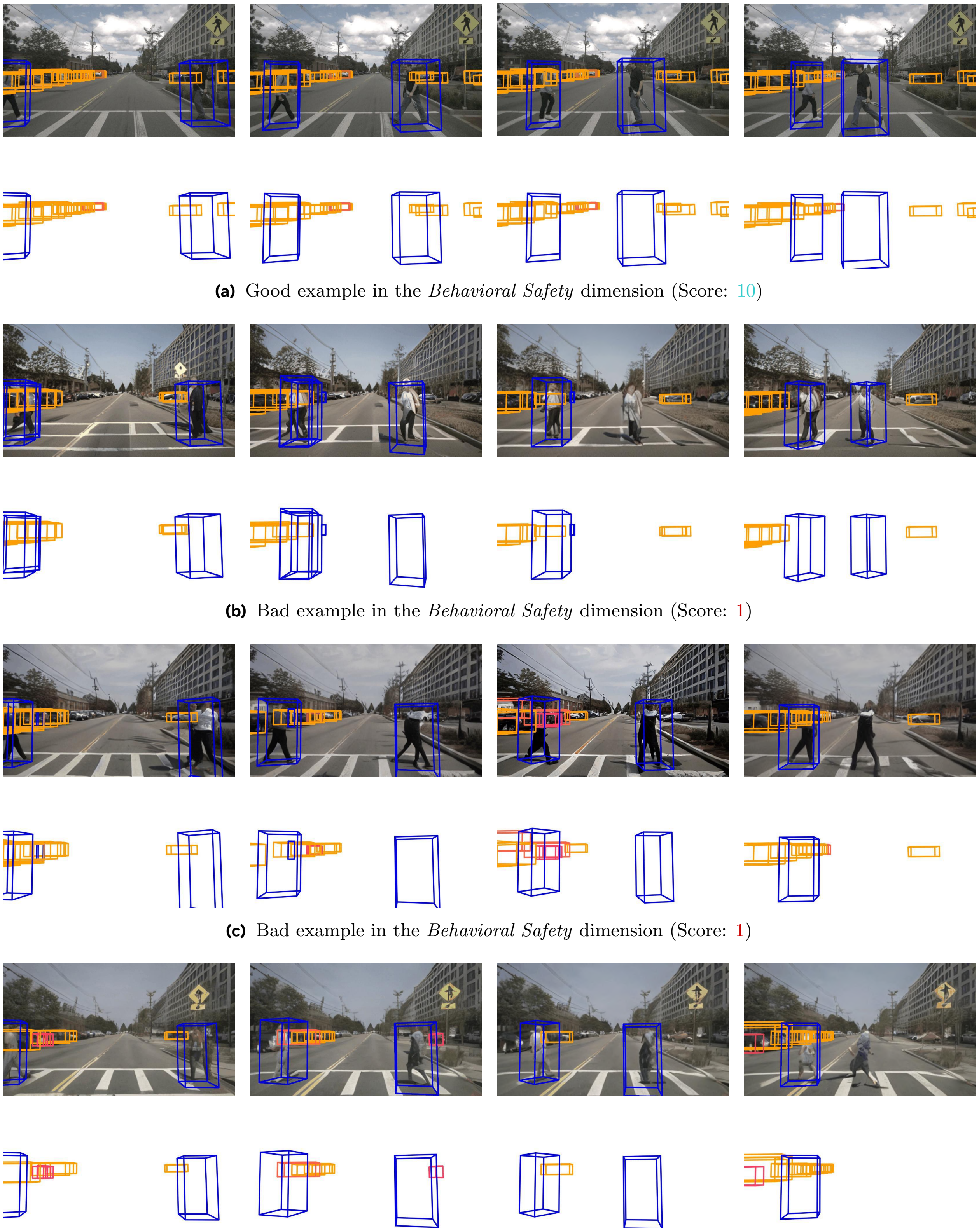
Figure 34 provides typical examples of videos with good and bad quality in terms of *Behavioral Safety*.

11.6.3 Evaluation & Analysis

Table 29 provides the complete results of models in terms of *Behavioral Safety*.

Table 29 Complete comparisons of state-of-the-art driving world models in terms of *Behavioral Safety* in WorldLens.

Behavioral Safety	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
min	2.0	2.0	2.0	2.0	2.0	2.0	-
max	4.0	4.0	10.0	8.0	6.0	4.0	-
mean	2.306	2.290	2.533	2.598	2.591	2.318	10
std	0.649	0.621	1.184	1.247	1.341	0.686	-
median	2.0	2.0	2.0	2.0	2.0	2.0	-
q25	2.0	2.0	2.0	2.0	2.0	2.0	-
q75	2.0	2.0	2.0	3.0	2.0	2.0	-



(a) Good example in the *Behavioral Safety* dimension (Score: 10)

(b) Bad example in the *Behavioral Safety* dimension (Score: 1)

(c) Bad example in the *Behavioral Safety* dimension (Score: 1)

(d) Bad example in the *Behavioral Safety* dimension (Score: 1)

Figure 34 Examples of "good" and "bad" human preference alignments in terms of *Behavioral Safety* in WorldLens.

12 Evaluation Agent

In this section, we present additional detail on the proposed **WorldLens-Agent** model, describing the architecture, prompting scheme, training setup, and providing some qualitative evaluation examples on out-of-distribution driving videos. We observe that evaluating generated worlds often hinges on human-centered criteria (physical plausibility) and subjective preferences (perceived realism) that quantitative metrics inherently miss. Our goal here is to train an auto-evaluation agent that can be utilized in a broader range of generated videos, and, simultaneously, align with the preferences of human annotators.

12.1 Agent Architecture

The **WorldLens-Agent** is a vision-language critic built on Qwen3-VL-8B [2] and trained to evaluate generated videos along human-centered dimensions, including overall realism, 3D consistency, physical plausibility, and behavioral safety.

As shown in Figure 35, the agent takes two types of input: an instruction text describing the evaluation criteria, which is processed by the frozen Qwen3 tokenizer, and a video generated by world models, which is encoded by the frozen Qwen3-VL vision encoder. The resulting features are projected into the language token space, forming a unified multimodal token sequence.

This sequence is then passed to the Qwen3-VL decoder, where LoRA adapters are applied only to the attention layers. All other components, including the vision encoder, the projector, the embedding layers, and the MLP blocks, remain frozen. This lightweight adaptation allows the model to incorporate human perceptual and safety-related priors learned from **WorldLens-26K**, enabling it to capture cues such as lighting realism, depth stability, object dynamics, and safety-critical violations while preserving the general multimodal capability of the base model.

Finally, the agent autoregressively produces a *structured JSON output* that contains a numerical score (1-10) and a concise rationale for each evaluation dimension. This representation yields a reliable, interpretable, and scalable assessment signal that complements conventional quantitative metrics and serves as a consistent preference oracle for world-model benchmarking and downstream reinforcement learning pipelines.

12.2 Prompt Scheme

The following prompting scheme specifies the instruction protocol for the WorldLens Evaluation Agent. Given a generated driving clip and a dimension-specific human rating rubric, the agent is guided to produce structured, evidence-based scores for multiple aspects of generative video quality.

The prompt enforces strict output formatting, dimension-aware reasoning, and rubric-consistent interpretation, ensuring reliable and reproducible automatic scoring.

12.3 Training Setup

The WorldLens-Agent is fine-tuned from Qwen3-VL-8B through supervised instruction tuning, allowing the model to better align with human evaluation preferences. We adopt LoRA adaptation on all attention modules, using a rank of 16 and a dropout rate of 0.05, which provides efficient preference learning while preserving the multimodal reasoning capabilities of the base model.

Training is performed for three epochs with a learning rate of 1e-4, cosine decay scheduling, and a warmup ratio of 0.1. All experiments are conducted on eight A100 GPUs using `bfloat16` precision. This configuration ensures stable convergence and effective adaptation, resulting in a vision-language critic that consistently captures realism, geometric consistency, physical plausibility, and safety-related cues in generated videos.

12.4 Qualitative Assessment

Figure 36 and Figure 37 present additional qualitative evaluations produced by the **WorldLens-Agent** on challenging driving scenarios, including **out-of-distribution** videos rendered or produced by Gen3C [83], Cosmos-Drive [82], and the CARLA [23] simulator. These examples illustrate the agent's ability to generalize

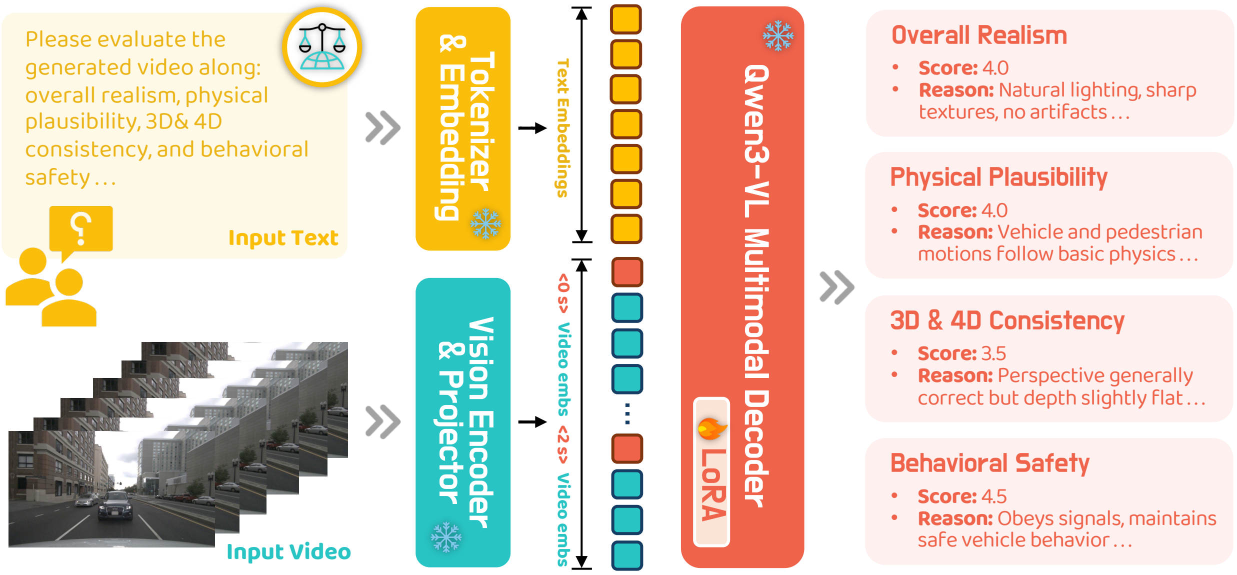

Please evaluate the generated video along: overall realism, physical plausibility, 3D& 4D consistency, and behavioral safety…
Input Text
Input Video
Tokenizer & Embedding
Vision Encoder & Projector
Text Embeddings
<0 s> Video embs <2 s> Video embs
Qwen3-VL Multimodal Decoder
LoRA
Overall Realism
Score: 4.0
Reason: Natural lighting, sharp textures, no artifacts…
Physical Plausibility
Score: 4.0
Reason: Vehicle and pedestrian motions follow basic physics…
3D & 4D Consistency
Score: 3.5
Reason: Perspective generally correct but depth slightly flat…
Behavioral Safety
Score: 4.5
Reason: Obeys signals, maintains safe vehicle behavior…

Figure 35 The architecture of the proposed **WorldLens-Agent** for auto-evaluation of generated driving videos.

beyond its training distribution and maintain consistent, human-aligned judgment across a wide spectrum of visual styles, scene structures, and motion dynamics.

As shown in Figure 36, the agent reliably identifies a broad range of safety-critical issues. These include *lane incursions*, *ignoring red lights*, and *near-collision events*, each accompanied by a concise explanation grounded in visible evidence. It also detects failures in physical plausibility, such as *unnatural animal motion* or vehicles exhibiting *incoherent dynamics*, where motion lacks realistic articulation or violates expected mass-gravity relationships. In the realm of realism, the agent highlights artifacts like *low-fidelity textures*, *simplified geometry*, and *game-engine rendering effects*, all of which degrade perceptual authenticity.

Figure 37 further demonstrates the agent's sensitivity to more severe and uncommon failure modes. It flags *physically impossible* ego-vehicle trajectories, such as the viewpoint unexpectedly lifting off the ground, as violations of basic mechanical and gravitational constraints. The agent also captures high-impact behavioral safety failures, including *colliding with stationary obstacles* such as ambulances or trucks. Beyond temporal or behavioral issues, it robustly identifies large-scale *3D and 4D consistency violations*, where buildings, vehicles, and road structures visibly intersect or pass through one another, indicating broken geometry and disrupted spatial coherence.

Together, these qualitative cases highlight the strong generalization capability of the proposed **WorldLens-Agent** and our ability to diagnose diverse, complex failure patterns across unseen generative video domains. The agent not only assigns scalar scores but also provides clear, interpretable rationales, enabling transparent and human-aligned evaluation under significant distribution shift.

System prompt for assessing the quality of generated driving videos with VLM

You are an expert **vision-language evaluator** designed to assess the quality of **AI-generated driving videos**. Given a short video clip together with a human-designed rubric, your task is to assign an objective score and provide a concise, evidence-based explanation. Your analysis focuses on a **single target dimension** at a time (*e.g.*, *overall_ realism*, *vehicle_ realism*, *pedestrian_ realism*, *3D_ - consistency*, *physical_ plausibility*, *behavioral_ safety*).

- **Evaluation Target**: The input specifies which dimension to evaluate, *e.g.*, `overall_realism`, `vehicle_realism`, etc.
- **Rubric-Guided Judgment**: Use the full English rubric provided for the selected dimension, strictly following its definitions, criteria, and scoring scale.
- **Evidence-Based Scoring**: Judge only what is **visually observable** in the video, such as temporal stability, geometry, textures, reflections, occlusions, physical consistency, and artifacts.
- **Score Range**: Output a numeric score in $[1, 10]$ with a step of 0.5 (i.e., $\{1.0, 1.5, \ldots, 10.0\}$), rounded to the nearest 0.5 and clamped to $[1, 10]$.
- **Rationale**: Provide a short English rationale that cites concrete visual evidence, *e.g.*, flicker, ghosting, shape distortions, interpenetration, lighting jumps, unsafe maneuvers, etc.

Scoring Rubrics for Each Dimension (Summary):

- **Overall Realism**
  - **1 – Highly Unrealistic:** Severe artifacts (warping, collapsing geometry, heavy flicker, broken roads, impossible lighting).
  - **3 – Unrealistic:** Structures roughly recognizable but textures blurry; perspective errors; frequent artifacts and instability.
  - **5 – Fair:** Mostly acceptable but clearly synthetic; noticeable unnatural boundaries, lighting jumps, or texture flicker.
  - **7 – Realistic:** Largely natural appearance; coherent lighting and geometry; only minor localized flaws.
  - **9 – Highly Realistic:** Almost indistinguishable from real dashcam footage; stable textures, correct perspective, consistent lighting.
- **Vehicle Realism**
  - **1 – Highly Unrealistic:** Strong distortions, split bodies, stretched parts, heavy flicker; vehicles look clearly fake.
  - **3 – Unrealistic:** Coarse material appearance, unstable reflections, poor temporal consistency, visible geometry defects.
  - **5 – Fair:** Car-like but with noticeable flaws in edges, contours, or materials; moderate instability across frames.
  - **7 – Realistic:** Mostly correct geometry, paint, glass, tires, and reflections; minor issues only.
  - **9 – Highly Realistic:** Faithful car shape and materials; natural glass/paint reflections; stable and coherent over time.
- **Pedestrian Realism**
  - **1 – Highly Unrealistic:** Broken limbs, twisted joints, ghosting, severe flicker, or missing body parts.
  - **3 – Unrealistic:** Human-like but coarse textures, inconsistent appearance, unnatural gait, or floating/sliding.
  - **5 – Fair:** Generally human-shaped with some artifacts or slightly rigid/unnatural motion.
  - **7 – Realistic:** Mostly natural appearance and motion; minor local inconsistencies acceptable.
  - **9 – Highly Realistic:** Convincing human geometry, clothing, motion, and lighting; temporally stable with no major artifacts.
- **3D Consistency**
  - **1 – Highly Inconsistent:** Strong jitter, scale popping, drift, disappear/reappear, incorrect depth ordering.
  - **3 – Inconsistent:** Overall motion roughly matches but unstable; noticeable mismatches or frequent small jumps.
  - **5 – Fair:** Broadly consistent with occasional jitters, mild drift, or minor alignment errors.
  - **7 – Consistent:** Mostly stable geometry and depth; only small irregularities.
  - **9 – Highly Consistent:** Smooth temporal evolution; stable shapes, positions, and trajectories with near-perfect depth coherence.
- **Physical Plausibility**
  - **1 – Highly Implausible:** Teleporting, merging/splitting objects, interpenetration, impossible shadows, large occlusion errors.
  - **3 – Implausible:** Noticeable but localized physics violations (*e.g.*, single intersection event, abrupt illumination jump).
  - **5 – Fair:** Mostly plausible but with several visible inconsistencies that do not dominate the clip.
  - **7 – Plausible:** Stable motion, contacts, and occlusions; only minor physical deviations.
  - **9 – Highly Plausible:** Fully consistent with real-world physics; smooth motion, proper occlusions, no interpenetration.
- **Behavioral Safety**
  - **1 – Highly Unsafe:** Impossible or dangerous behaviors; collisions, severe occlusion failures preventing safety judgment.
  - **3 – Unsafe:** Localized unsafe interactions, obvious violations of signals or right-of-way; unstable visuals harming safety perception.
  - **5 – Generally Safe:** Mostly reasonable behaviors with small issues that do not fundamentally affect safety.
  - **7 – Safe:** Stable, predictable motion; compliant with visible cues; minor cosmetic defects only.
  - **9 – Highly Safe:** Clear, unambiguous, low-risk behavior; natural interaction patterns consistent with real traffic rules.

Instruction:

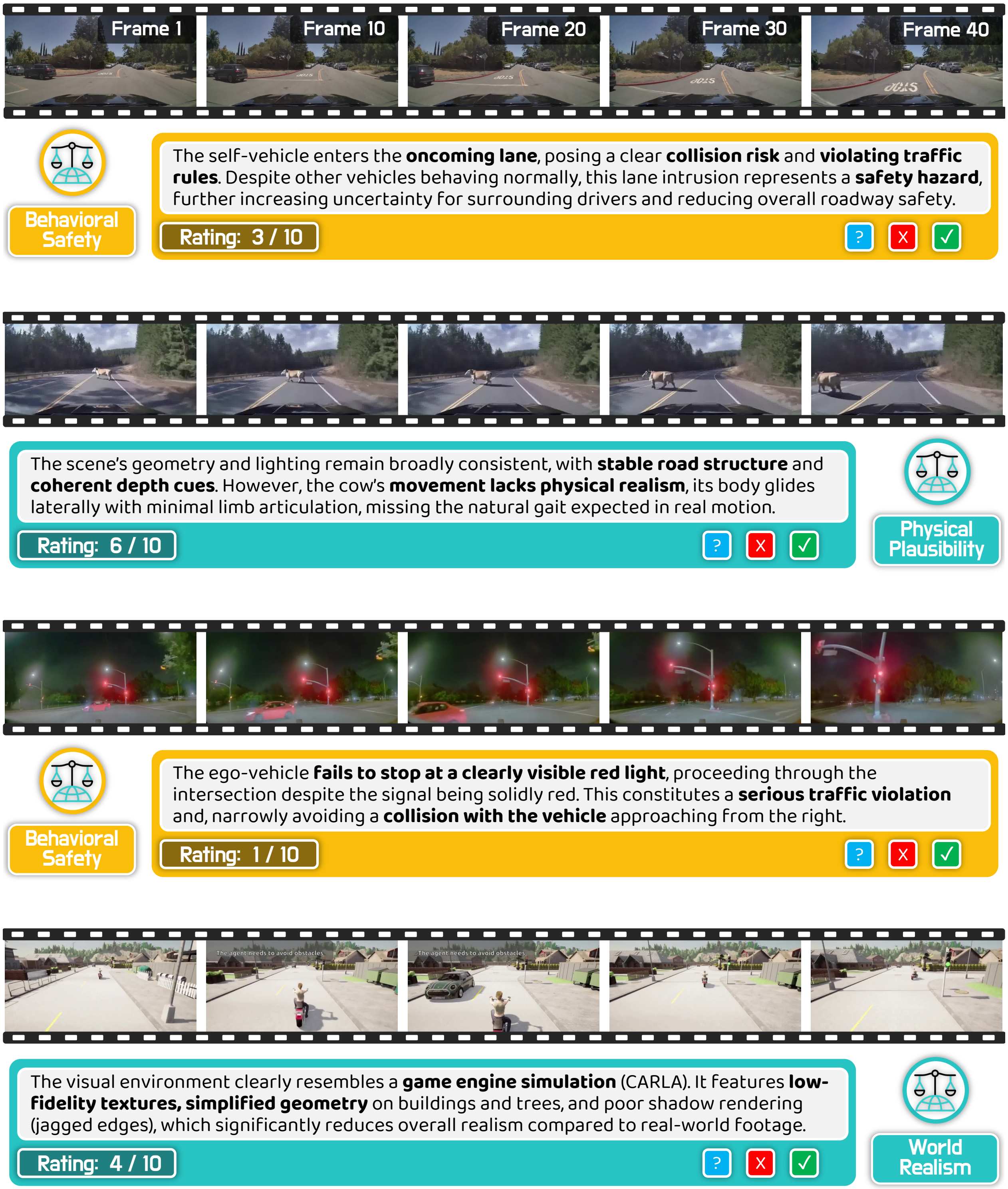
- Evaluate the clip **only** from visible evidence; do not hallucinate or infer information not shown in the video.
- Follow the corresponding rubric for the given dimension when mapping visual quality to a score in $[1, 10]$.
- The output **must** be a single valid JSON object with exactly two keys: `"score"` and `"reason"`.
- Do **not** include any extra keys, text, Markdown, or comments outside the JSON object.

Please format your results as follows:

```
{
  "score": <one of 1.0,1.5,...,10.0>
  "reason": "<Brief, evidence-based explanation based only on visible cues.>"
}
```

Example output:

```
{
  "score": 2.5,
  "reason": "Frequent texture flicker on vehicles and unstable shadows reduce realism, but
      geometry and traffic behaviors remain mostly plausible."
}
```



The self-vehicle enters the **oncoming lane**, posing a clear **collision risk** and **violating traffic rules**. Despite other vehicles behaving normally, this lane intrusion represents a **safety hazard**, further increasing uncertainty for surrounding drivers and reducing overall roadway safety.

Rating: 3 / 10 ? X ✓

The scene's geometry and lighting remain broadly consistent, with **stable road structure** and **coherent depth cues**. However, the cow's **movement lacks physical realism**, its body glides laterally with minimal limb articulation, missing the natural gait expected in real motion.

Rating: 6 / 10 ? X ✓

The ego-vehicle **fails to stop at a clearly visible red light**, proceeding through the intersection despite the signal being solidly red. This constitutes a **serious traffic violation** and, narrowly avoiding a **collision with the vehicle** approaching from the right.

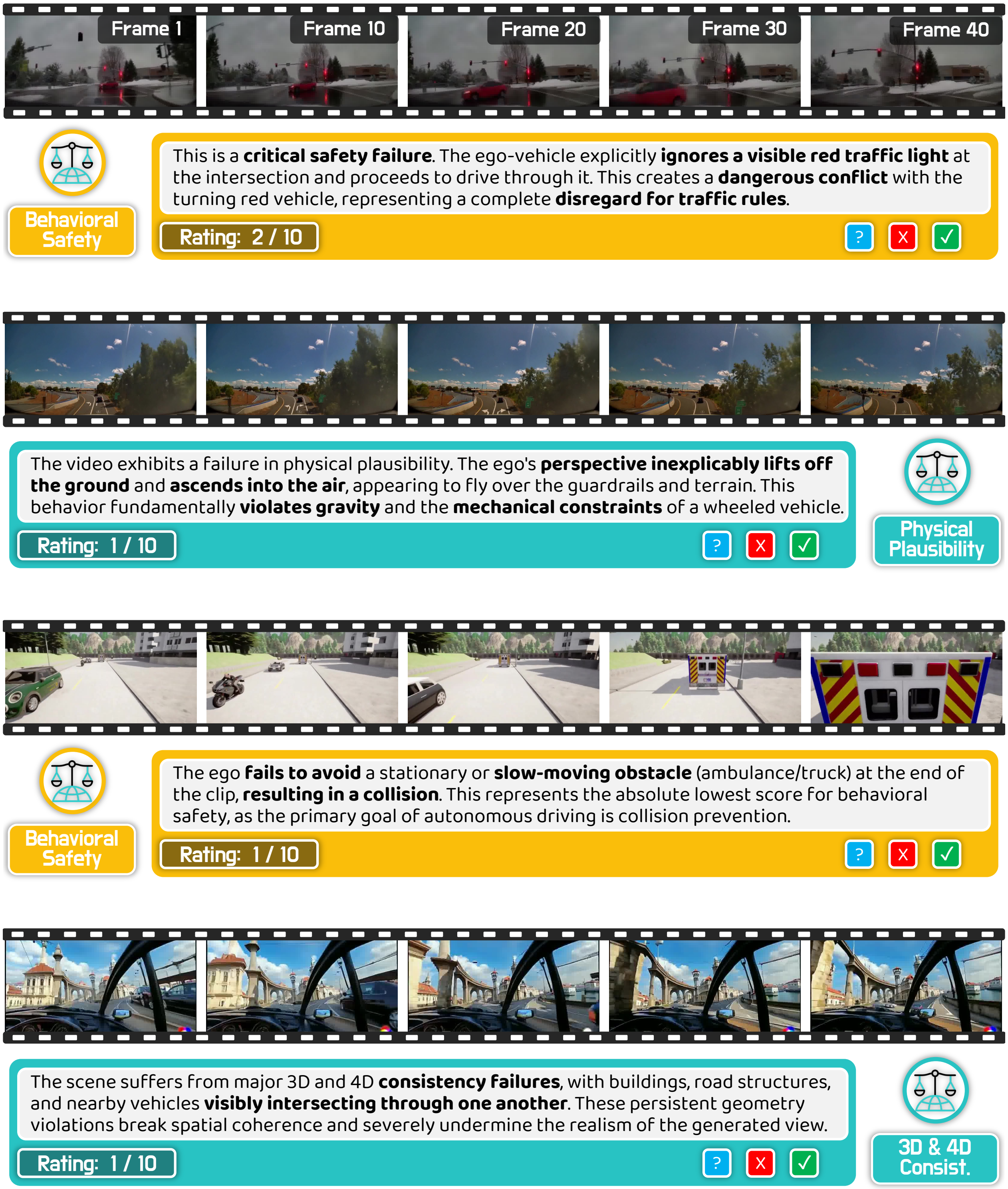
Rating: 1 / 10 ? X ✓

The agent needs to avoid obstacles
The agent needs to avoid obstacles

The visual environment clearly resembles a **game engine simulation** (CARLA). It features **low-fidelity textures, simplified geometry** on buildings and trees, and poor shadow rendering (jagged edges), which significantly reduces overall realism compared to real-world footage.

Rating: 4 / 10 ? X ✓

Figure 36 Additional qualitative assessments of the **WorldLens-Agent** evaluation on challenging driving conditions.



This is a **critical safety failure**. The ego-vehicle explicitly **ignores a visible red traffic light** at the intersection and proceeds to drive through it. This creates a **dangerous conflict** with the turning red vehicle, representing a complete **disregard for traffic rules**.

Rating: 2 / 10

The video exhibits a failure in physical plausibility. The ego's **perspective inexplicably lifts off the ground** and **ascends into the air**, appearing to fly over the guardrails and terrain. This behavior fundamentally **violates gravity** and the **mechanical constraints** of a wheeled vehicle.

Rating: 1 / 10

The ego **fails to avoid** a stationary or **slow-moving obstacle** (ambulance/truck) at the end of the clip, **resulting in a collision**. This represents the absolute lowest score for behavioral safety, as the primary goal of autonomous driving is collision prevention.

Rating: 1 / 10

The scene suffers from major 3D and 4D **consistency failures**, with buildings, road structures, and nearby vehicles **visibly intersecting through one another**. These persistent geometry violations break spatial coherence and severely undermine the realism of the generated view.

Rating: 1 / 10

Figure 37 Additional qualitative assessments of the **WorldLens-Agent** evaluation on challenging driving conditions.

13 Broader Impact & Limitations

In this section, we elaborate on the broader impact, societal influence, and potential limitations of the proposed approach.

13.1 Broader Impact

Our benchmark advances the evaluation of generative world models by establishing a unified, transparent, and reproducible protocol that links perception, geometry, physics, and behavior. By grounding quantitative scores in human perception and physical reasoning, we encourage the development of models that are not only visually convincing but also physically reliable and functionally safe.

The benchmark, dataset, and agent together promote standardization and comparability in this rapidly evolving domain, helping researchers diagnose weaknesses, track progress, and design more robust embodied simulators. Beyond autonomous driving, the framework can inspire principled evaluation methods for robotics, AR/VR simulation, and broader world-model research.

13.2 Societal Influence

Our benchmark has implications for AI safety, trustworthy simulation, and embodied intelligence. By providing quantitative and human-aligned metrics for realism, physical plausibility, and behavioral safety, our benchmark helps mitigate risks from models that may appear realistic but behave unrealistically when used for planning or training downstream agents. Reliable evaluation of generative simulators could accelerate applications in safe-driving research, synthetic dataset generation, and policy testing under controlled conditions.

Nonetheless, the framework should be used responsibly, especially when synthetic data influence safety-critical decisions, ensuring transparency in evaluation and avoiding misuse for deceptive content generation.

13.3 Potential Limitations

While our benchmark provides a comprehensive evaluation spectrum, several limitations remain. First, the benchmark currently focuses on driving-world scenarios; extending to indoor, aerial, or humanoid environments requires additional metrics and domain-specific cues. Second, although the human preference dataset (WorldLens-26K) captures rich perceptual reasoning, it may reflect annotator bias toward specific visual styles or regions, which future work could mitigate through more diverse and cross-cultural labeling. Third, the evaluation agent, though effective in zero-shot settings, inherits limitations from its underlying language model and supervision quality. Lastly, physical realism in simulation is inherently open-ended; new metrics may be required as models evolve toward interactive and multimodal 4D reasoning.

14 Public Resource Used

In this section, we acknowledge the use of the following public resources, during the course of this work:

• nuScenes[1]	CC BY-NC-SA 4.0
• nuscenes-devkit[2]	Apache License 2.0
• KITTI[3]	Non-Commercial Use Only (Research Purposes)
• waymo-open-dataset[4]	Apache License 2.0
• MagicDrive[5]	Apache License 2.0

[1] https://www.nuscenes.org/nuscenes.

[2] https://github.com/nutonomy/nuscenes-devkit.

[3] http://www.cvlibs.net/datasets/kitti.

[4] https://github.com/waymo-research/waymo-open-dataset.

[5] https://github.com/cure-lab/MagicDrive.

DreamForge[6]	Apache License 2.0
DriveDreamer-2[7]	Apache License 2.0
OpenDWM[8]	MIT License
DiST-4D[9]	None
$\mathcal{X}$-Scene[10]	None
Panacea[11]	Apache License 2.0
Limsim[12]	None
DriveStudio[13]	MIT License
DriveArena[14]	None
DrivingSphere[15]	Apache License 2.0
MagicDrive-V2[16]	AGPL-3.0 license
UniAD[17]	Apache License 2.0
Open3D[18]	MIT License
PyTorch[19]	BSD License
ROS Humble[20]	Apache License 2.0
torchsparse[21]	MIT License
VBench[22]	Apache License 2.0
SparseOcc[23]	Apache License 2.0
DINO[24]	Apache License 2.0
DINOv2[25]	Apache License 2.0
MMEngine[26]	Apache License 2.0
MMCV[27]	Apache License 2.0
MMDetection[28]	Apache License 2.0
MMDetection3D[29]	Apache License 2.0

[6] https://github.com/PJLab-ADG/DriveArena.
[7] https://github.com/f1yfisher/DriveDreamer2.
[8] https://github.com/SenseTime-FVG/OpenDWM.
[9] https://github.com/royalmelon0505/dist4d.
[10] https://github.com/yuyang-cloud/X-Scene.
[11] https://github.com/wenyuqing/panacea.
[12] https://github.com/PJLab-ADG/LimSim/tree/LimSim_plus.
[13] https://github.com/ziyc/drivestudio.
[14] https://github.com/PJLab-ADG/DriveArena.
[15] https://github.com/yanty123/DrivingSphere.
[16] https://github.com/flymin/MagicDrive-V2.
[17] https://github.com/OpenDriveLab/UniAD.
[18] http://www.open3d.org.
[19] https://pytorch.org.
[20] https://docs.ros.org/en/humble.
[21] https://github.com/mit-han-lab/torchsparse.
[22] https://github.com/Vchitect/VBench.
[23] https://github.com/MCG-NJU/SparseOcc.
[24] https://github.com/facebookresearch/dino.
[25] https://github.com/facebookresearch/dinov2.
[26] https://github.com/open-mmlab/mmengine.
[27] https://github.com/open-mmlab/mmcv.
[28] https://github.com/open-mmlab/mmdetection.
[29] https://github.com/open-mmlab/mmdetection3d.

- OpenPCSeg[30] . . . . Apache License 2.0
- OpenPCDet[31] . . . . Apache License 2.0
- Qwen3-VL[32] . . . . Apache License 2.0
- LLaMA-Factory[33] . . . . Apache License 2.0

References

[1] Hidehisa Arai, Keishi Ishihara, Tsubasa Takahashi, and Yu Yamaguchi. ACT-Bench: Towards action controllable world models for autonomous driving. *arXiv preprint arXiv:2412.05337*, 2024.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models, 2025.

[4] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia*, pages 1–11, 2024.

[5] Florent Bartoccioni, Elias Ramzi, Victor Besnier, Shashanka Venkataramanan, Tuan-Hung Vu, Yihong Xu, Loick Chambon, Spyros Gidaris, Serkan Odabas, David Hurych, Renaud Marlet, Alexandre Boulch, Mickael Chen, Éloi Zablocki, Andrei Bursuc, Eduardo Valle, and Matthieu Cord. VaViM and VaVAM: Autonomous driving through video generative modeling. *arXiv preprint arXiv:2502.15672*, 2025.

[6] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*, 2(3):8, 2023.

[7] Hengwei Bian, Lingdong Kong, Haozhe Xie, Liang Pan, Yu Qiao, and Ziwei Liu. DynamicCity: Large-scale 4D occupancy generation from dynamic scenes. In *International Conference on Learning Representations*, 2025.

[8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[9] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[10] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

[30] https://github.com/PJLab-ADG/OpenPCSeg.

[31] https://github.com/open-mmlab/OpenPCDet.

[32] https://github.com/QwenLM/Qwen3-VL.

[33] https://github.com/hiyouga/LLaMA-Factory.

[11] Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. In *Conference on Robot Learning*. PMLR, 2025.

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[13] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[14] Anthony Chen, Wenzhao Zheng, Yida Wang, Xueyang Zhang, Kun Zhan, Peng Jia, Kurt Keutzer, and Shanghang Zhang. GeoDrive: 3D geometry-informed driving world model with precise action control. *arXiv preprint arXiv:2505.22421*, 2025.

[15] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22831–22840, 2025.

[16] Yuntao Chen, Yuqi Wang, and Zhaoxiang Zhang. DrivingGPT: Unifying driving world modeling and planning with multi-modal autoregressive transformers. *arXiv preprint arXiv:2412.18607*, 2024.

[17] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. OmniRe: Omni urban scene reconstruction. In *International Conference on Learning Representations*, 2025.

[18] Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025.

[19] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[20] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems*, volume 37, pages 28706–28719, 2024.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[22] Shuxiao Ding, Lukas Schneider, Marius Cordts, and Juergen Gall. ADA-Track: End-to-end multi-camera 3D multi-object tracking with alternating detection and association. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15184–15194, 2024.

[23] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Conference on Robot Learning*, pages 1–16. PMLR, 2017.

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[25] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025.

[26] Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, et al. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. *arXiv preprint arXiv:2501.08453*, 2025.

[27] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. MagicDrive: Street view generation with diverse 3D geometry control. In *International Conference on Learning Representations*, 2023.

[28] Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. MagicDrive-V2: High-resolution long video generation for autonomous driving with adaptive control. In *IEEE/CVF International Conference on Computer Vision*, pages 28135–28144, 2025.

[29] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[30] Jiazhe Guo, Yikang Ding, Xiwu Chen, Shuo Chen, Bohan Li, Yingshuang Zou, Xiaoyang Lyu, Feiyang Tan, Xiaojuan Qi, Zhiheng Li, and Hao Zhao. DiST-4D: Disentangled spatiotemporal diffusion with metric depth for 4D driving scene generation. In *IEEE/CVF International Conference on Computer Vision*, pages 27231–27241, 2025.

[31] Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. MineWorld: A real-time and open-source interactive world model on MineCraft. *arXiv preprint arXiv:2504.08388*, 2025.

[32] Mariam Hassan, Sebastian Stapf, Ahmad Rahimi, Pedro M B Rezende, Yasaman Haghighi, David Brüggemann, Isinsu Katircioglu, Lin Zhang, Xiaoran Chen, Suman Saha, Marco Cannici, Elie Aljalbout, Botao Ye, Xi Wang, Aram Davtyan, Mathieu Salzmann, Davide Scaramuzza, Marc Pollefeys, Paolo Favaro, and Alexandre Alahi. GEM: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22404–22415, 2025.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[34] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based object re-identification. In *IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021.

[35] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two-time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30:6629–6640, 2017.

[36] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[37] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[38] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023.

[39] Binyuan Huang, Yuqing Wen, Yucheng Zhao, Yaosi Hu, Yingfei Liu, Fan Jia, Weixin Mao, Tiancai Wang, Chi Zhang, Chang Wen Chen, Zhenzhong Chen, and Xiangyu Zhang. SubjectDrive: Scaling generative data in autonomous driving via subject control. In *AAAI Conference on Artificial Intelligence*, volume 39, pages 3617–3625, 2025.

[40] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

[41] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[42] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. VBench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024.

[43] Ziqi Huang, Ning Yu, Gordon Chen, Haonan Qiu, Paul Debevec, and Ziwei Liu. VChain: Chain-of-visual-thought for reasoning in video generation. *arXiv preprint arXiv:2510.05094*, 2025.

[44] Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. ADriver-I: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023.

[45] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: Vectorized scene representation for efficient autonomous driving. In *IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023.

[46] Junpeng Jiang, Gangyi Hong, Lijun Zhou, Enhui Ma, Hengtong Hu, Xia Zhou, Jie Xiang, Fan Liu, Kaicheng Yu, Haiyang Sun, Kun Zhan, Peng Jia, and Miao Zhang. DiVE: DiT-based video generation with enhanced control. *arXiv preprint arXiv:2409.01595*, 2024.

[47] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

[48] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[49] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021.

[50] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

[51] Akio Kodaira, Chenfeng Xu, Toshiki Hazama, Takanori Yoshimoto, Kohei Ohno, Shogo Mitsuhori, Soichi Sugano, Hanying Cho, Zhijian Liu, Masayoshi Tomizuka, et al. StreamDiffusion: A pipeline-level solution for real-time interactive generation. In *IEEE/CVF International Conference on Computer Vision*, pages 12371–12380, 2025.

[52] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Yaru Niu, Wei Tsang Ooi, Benoit R. Cottereau, Lai Xing Ng, Yuexin Ma, Wenwei Zhang, Liang Pan, Kai Chen, Ziwei Liu, Weichao Qiu, Wei Zhang, Xu Cao, Hao Lu, Ying-Cong Chen, et al. The RoboDrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816*, 2024.

[53] Lingdong Kong, Wesley Yang, Jianbiao Mei, Youquan Liu, Ao Liang, Dekai Zhu, Dongyue Lu, Wei Yin, Xiaotao Hu, Mingkai Jia, Junyuan Deng, Kaiwen Zhang, Yang Wu, Tianyi Yan, Shenyuan Gao, Song Wang, Linfeng Li, Liang Pan, Yong Liu, Jianke Zhu, Wei Tsang Ooi, Steven C. H. Hoi, and Ziwei Liu. 3D and 4D world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025.

[54] Pika Lab. Pika. https://www.pika.art, 2024.

[55] Chengxuan Li, Di Huang, Zeyu Lu, Yang Xiao, Qingqi Pei, and Lei Bai. A survey on long video generation: Challenges, methods, and prospects. *arXiv preprint arXiv:2403.16407*, 2024.

[56] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. DrivingDiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model. In *European Conference on Computer Vision*, pages 469–485. Springer, 2024.

[57] Zhikai Li, Xuewen Liu, Dongrong Joe Fu, Jianquan Li, Qingyi Gu, Kurt Keutzer, and Zhen Dong. K-sort arena: Efficient and reliable benchmarking for generative models via k-wise human preferences. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9131–9141, 2025.

[58] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from LiDAR-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2025.

[59] Ao Liang, Lingdong Kong, Dongyue Lu, Youquan Liu, Jian Fang, Huaici Zhao, and Wei Tsang Ooi. Perspective-invariant 3D object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 27725–27738, 2025.

[60] Ao Liang, Youquan Liu, Yu Yang, Dongyue Lu, Linfeng Li, Lingdong Kong, Huaici Zhao, and Wei Tsang Ooi. LiDARCrafter: Dynamic 4D world modeling from LiDAR sequences. In *AAAI Conference on Artificial Intelligence*, volume 40, 2026.

[61] Mingxiang Liao, Qixiang Ye, Wangmeng Zuo, Fang Wan, Tianyu Wang, Yuzhong Zhao, Jingdong Wang, and Xinyu Zhang. Evaluation of text-to-video generation models: A dynamics perspective. In *Advances in Neural Information Processing Systems*, volume 37, pages 109790–109816, 2024.

[62] Hongbin Lin, Zilu Guo, Yifan Zhang, Shuaicheng Niu, Yafeng Li, Ruimao Zhang, Shuguang Cui, and Zhen Li. DriveGen: Generalized and robust 3D detection in driving via controllable text-to-image diffusion generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27497–27507, 2025.

[63] Xiao Liu, Xinhao Xiang, Zizhong Li, Yongheng Wang, Zhuoheng Li, Zhuosheng Liu, Weidi Zhang, Weiqi Ye, and Jiawei Zhang. A survey of AI-generated video evaluation. *arXiv preprint arXiv:2410.19884*, 2024.

[64] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024.

[65] Youquan Liu, Lingdong Kong, Weidong Yang, Xin Li, Ao Liang, Runnan Chen, Ben Fei, and Tongliang Liu. La La LiDAR: Large-scale layout generation from LiDAR data. In *AAAI Conference on Artificial Intelligence*, volume 40, 2026.

[66] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L. Rus, and Song Han. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation*, pages 2774–2781, 2023.

[67] Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. WoVoGen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *European Conference on Computer Vision*, pages 329–345. Springer, 2024.

[68] Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Xuanhua He, Chenyang Zhu, Hongyu Liu, Yingqing He, Zeyu Wang, Zhifeng Li, Xiu Li, Wei Liu, Dan Xu, Linfeng Zhang, and Qifeng Chen. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025.

[69] Jianbiao Mei, Tao Hu, Xuemeng Yang, Licheng Wen, Yu Yang, Tiantian Wei, Yukai Ma, Min Dou, Botian Shi, and Yong Liu. DreamForge: Motion-aware autoregressive video generation for multi-view driving scenes. *arXiv preprint arXiv:2409.04003*, 2024.

[70] Jianbiao Mei, Yu Yang, Xuemeng Yang, Licheng Wen, Jiajun Lv, Botian Shi, and Yong Liu. Vision-centric 4d occupancy forecasting and planning via implicit residual world models. *arXiv preprint arXiv:2510.16729*, 2025.

[71] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106, 2021.

[72] Jingcheng Ni, Yuxin Guo, Yichen Liu, Rui Chen, Lewei Lu, and Zehuan Wu. MaskGWM: A generalizable driving world model with video mask reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22381–22391, 2025.

[73] On-Road Automated Driving (ORAD) Committee. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. https://www.sae.org/standards/j3016_202104-taxonomy-definitions-terms-related-driving-automation-systems-road-motor-vehicles, 2021.

[74] OpenAI. Sora. Accessed February 15, 2024 [Online] https://sora.com/library, 2024. URL https://sora.com/library.

[75] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

[76] Gary Overett, Lars Petersson, Nathan Brewer, Lars Andersson, and Niklas Pettersson. A new pedestrian dataset for supervised learning. In *IEEE Intelligent Vehicles Symposium*, pages 373–378, 2008.

[77] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model, 2024. URL https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/.

[78] Jordan Peper, Zhenjiang Mao, Yuang Geng, Siyuan Pan, and Ivan Ruchkin. Four principles for physically interpretable world models. *arXiv preprint arXiv:2503.02143*, 2025.

[79] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual

models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[80] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

[81] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[82] Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, Seung Wook Kim, Jun Gao, Laura Leal-Taixe, Mike Chen, Sanja Fidler, and Huan Ling. Cosmos-Drive-Dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025.

[83] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3C: 3D-informed world-consistent video generation with precise camera control. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6132, 2025.

[84] Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. VideoWorld: Exploring knowledge learning from unlabeled videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29029–29039, 2025.

[85] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[86] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. GAIA-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

[87] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[88] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. *Advances in Neural Information Processing Systems*, 29:2234–2242, 2016.

[89] SenseTime-FVG. Open Driving World Models (OpenDWM). https://github.com/SenseTime-FVG/OpenDWM, 2025.

[90] Chenyang Si, Weichen Fan, Zhengyao Lv, Ziqi Huang, Yu Qiao, and Ziwei Liu. RepVideo: Rethinking cross-layer representation for video generation. *arXiv preprint arXiv:2501.08994*, 2025.

[91] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6572–6582, 2024.

[92] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021.

[93] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8406–8416, 2025.

[94] Lei Sun, Kaiwei Wang, Kailun Yang, and Kaite Xiang. See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion. In *Artificial Intelligence and Machine Learning in Defense Applications*, volume 11169, pages 77–89. SPIE, 2019.

[95] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[96] Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. SparseOCC: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15035–15044, 2024.

[97] Google Team. Veo2. Accessed December 18, 2024 [Online] https://deepmind.google/technologies/veo/veo-2/, 2025. URL https://deepmind.google/technologies/veo/veo-2/.

[98] Kuaishou Team. Kling. Accessed December 9, 2024 [Online] https://klingai.kuaishou.com/, 2024. URL https://klingai.kuaishou.com/.

[99] Tecent Team. HunyuanVideo: A systematic framework for large video generative models, 2024.

[100] Wan Team. Wan: Open and advanced large-scale video generative models, 2025.

[101] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[102] Jiarui Wang, Juntong Wang, Xiaorong Zhu, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min. AIGVQA: A unified framework for multi-dimensional quality assessment of AI-generated video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 93383–3390, 2025.

[103] Xiaodong Wang and Peixi Peng. ProphetDWM: A driving world model for rolling out future actions and videos. *arXiv preprint arXiv:2505.18650*, 2025.

[104] Xiaodong Wang, Zhirong Wu, and Peixi Peng. LongDWM: Cross-granularity distillation for building a long-term driving world model. *arXiv preprint arXiv:2506.01546*, 2025.

[105] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024.

[106] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025.

[107] Yuping Wang, Shuo Xing, Cui Can, Renjie Li, Hongyuan Hua, Kexin Tian, Zhaobin Mo, Xiangbo Gao, Keshu Wu, Sulong Zhou, Hengxu You, Juntong Peng, Junge Zhang, Zehao Wang, Rui Song, Mingxuan Yan, Walter Zimmer, Xingcheng Zhou, Peiran Li, Zhaohan Lu, Chia-Ju Chen, Yue Huang, Ryan A. Rossi, Lichao Sun, Hongkai Yu, Zhiwen Fan, Frank Hao Yang, Yuhao Kang, Ross Greer, Chenxi Liu, Eun Hak Lee, Xuan Di, Xinyue Ye, Liu Ren, Alois Knoll, Xiaopeng Li, Shuiwang Ji, Masayoshi Tomizuka, Marco Pavone, Tianbao Yang, Jing Du, Ming-Hsuan Yang, Hua Wei, Ziran Wang, Yang Zhou, Jiachen Li, and Zhengzhong Tu. Generative AI for autonomous driving: Frontiers and opportunities. *arXiv preprint arXiv:2505.08854*, 2025.

[108] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024.

[109] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[110] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[111] Licheng Wen, Daocheng Fu, Song Mao, Pinlong Cai, Min Dou, Yikang Li, and Yu Qiao. LimSim: A long-term interactive multi-scenario traffic simulator. In *IEEE International Conference on Intelligent Transportation Systems*, pages 1255–1262, 2023.

[112] Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6902–6912, 2024.

[113] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020.

[114] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

[115] Wei Wu, Xi Guo, Weixuan Tang, Tingxuan Huang, Chiyu Wang, and Chenjing Ding. DriveScape: Towards high-resolution controllable multi-view driving video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17187–17196, 2025.

[116] Yanhao Wu, Haoyang Zhang, Tianwei Lin, Lichao Huang, Shujie Luo, Rui Wu, Congpei Qiu, Wei Ke, and Tong Zhang. Generating multimodal driving scenes via next-scene prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6844–6853, 2025.

[117] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090, 2021.

[118] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are VLMs ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. In *IEEE/CVF International Conference on Computer Vision*, pages 6585–6597, 2025.

[119] Haiwei Xue, Xiangyang Luo, Zhanghao Hu, Xin Zhang, Xunzhi Xiang, Yuqin Dai, Jianzhuang Liu, Zhensong Zhang, Minglei Li, Jian Yang, Fei Ma, Zhiyong Wu, Changpeng Yang, Zonghong Dai, and Fei Richard Yu. Human motion video generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):10709–10730, 2025.

[120] Tianyi Yan, Wencheng Han, Xia Zhou, Xueyang Zhang, Kun Zhan, Cheng-Zhong Xu, and Jianbing Shen. RLGF: Reinforcement learning with geometric feedback for autonomous driving video generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[121] Tianyi Yan, Dongming Wu, Wencheng Han, Junpeng Jiang, Xia Zhou, Kun Zhan, Cheng zhong Xu, and Jianbing Shen. DrivingSphere: Building a high-fidelity 4D world for closed-loop simulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27531–27541, 2025.

[122] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video generation using VQ-VAE and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[123] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *European Conference on Computer Vision*, pages 156–173. Springer, 2024.

[124] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized predictive model for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14662–14672, 2024.

[125] Kairui Yang, Enhui Ma, Jibin Peng, Qing Guo, Di Lin, and Kaicheng Yu. BEVControl: Accurately controlling street-view elements with multi-perspective consistency via BEV sketch layout. *arXiv preprint arXiv:2308.01661*, 2023.

[126] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, volume 37, pages 21875–21911, 2024.

[127] Xuemeng Yang, Licheng Wen, Yukai Ma, Jianbiao Mei, Xin Li, Tiantian Wei, Wenjie Lei, Daocheng Fu, Pinlong Cai, Min Dou, Botian Shi, Liang He, Yong Liu, and Yu Qiao. DriveArena: A closed-loop generative simulation platform for autonomous driving. In *IEEE/CVF International Conference on Computer Vision*, pages 26933–26943, 2025.

[128] Yu Yang, Alan Liang, Jianbiao Mei, Yukai Ma, Yong Liu, and Gim Hee Lee. X-Scene: Large-scale driving scene generation with high fidelity and flexible controllability. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[129] Yu Yang, Jianbiao Mei, Yukai Ma, Siliang Du, Wenqing Chen, Yijie Qian, Yuxiang Feng, and Yong Liu. Driving in the occupancy world: Vision-centric 4d occupancy forecasting and planning via world models for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9327–9335, 2025.

[130] Zhao Yang, Zezhong Qian, Xiaofan Li, Weixiang Xu, Gongpeng Zhao, Ruohong Yu, Lingsi Zhu, and Longjun Liu. DualDiff+: Dual-branch diffusion for high-fidelity video generation with reward guidance. *arXiv preprint arXiv:2503.03689*, 2025.

[131] Zhuoran Yang, Xi Guo, Chenjing Ding, Chiyu Wang, and Wei Wu. Physical informed driving world model. *arXiv preprint arXiv:2412.08410*, 2024.

[132] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[133] Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun Wang, Bernard Ghanem, Huchuan Lu, Chaochao Lu, Wanli Ouyang, Yu Qiao, Philip Torr, and Jing Shao. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024.

[134] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. GSplat: An open-source library for Gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025.

[135] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. WonderWorld: Interactive 3D scene generation from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5916–5926, 2025.

[136] Jingtong Yue, Ziqi Huang, Zhaoxi Chen, Xintao Wang, Pengfei Wan, and Ziwei Liu. Simulating the world model with artificial intelligence: A roadmap. *arXiv preprint arXiv:2511.08585*, 2025.

[137] Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. Evaluation agent: Efficient and promptable evaluation framework for visual generative models. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

[138] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023.

[139] Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, Xun Cao, and Wei Yin. Epona: Autoregressive diffusion world model for autonomous driving. In *IEEE/CVF International Conference on Computer Vision*, pages 27220–27230, 2025.

[140] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[141] Zhichao Zhang, Wei Sun, and Guangtao Zhai. A perspective on quality evaluation for AI-generated videos. *Sensors*, 25:4668, 2025.

[142] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. DriveDreamer-2: LLM-enhanced world models for diverse driving video generation. In *AAAI Conference on Artificial Intelligence*, volume 39, pages 10412–10420, 2025.

[143] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

[144] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE/CVF International Conference on Computer Vision*, pages 1116–1124, 2015.

[145] Xin Zhou, Dingkang Liang, Sifan Tu, Xiwu Chen, Yikang Ding, Dingyuan Zhang, Feiyang Tan, Hengshuang Zhao, and Xiang Bai. HERMES: A unified self-driving world model for simultaneous 3D scene understanding and generation. In *IEEE/CVF International Conference on Computer Vision*, pages 27817–27827, 2025.

[146] Yunsong Zhou, Michael Simon, Zhenghao Mark Peng, Sicheng Mo, Hongzi Zhu, Minyi Guo, and Bolei Zhou. SimGen: Simulator-conditioned driving scene generation. In *Advances in Neural Information Processing Systems*, volume 37, pages 48838–48874, 2024.

[147] Dekai Zhu, Yixuan Hu, Youquan Liu, Dongyue Lu, Lingdong Kong, and Slobodan Ilic. SPIRAL: Semantic-aware progressive LiDAR scene generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[148] Jialong Zuo, Ying Nie, Hanyu Zhou, Huaxin Zhang, Haoyu Wang, Tianyu Guo, Nong Sang, and Changxin Gao. Cross-video identity correlating for person re-identification pre-training. *Advances in Neural Information Processing Systems*, 37:25228–25250, 2024.