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

Supplementary Material

Table of Contents

A. 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.

A.1. Subject Fidelity

A.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.

A.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:

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

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 that align with real-world appearances, 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.

A.1.3. Implementation Details

We evaluate *Subject Fidelity* using class-specific confidence scores. Pedestrian crops are classified using a pedestrian classifier pretrained on several commonly used pedestrian-datasets [6, 18, 27, 35], while vehicle crops are classified with a ViT-B/16 model (`'google/vit-base-patch16-224'`) [29] pretrained on ImageNet-21k (14 million images, 21,843 classes) [8]. 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.

A.1.4. Examples

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

A.1.5. Evaluation & Analysis

Tab. I provides the complete results of models in terms of *Subject Fidelity*.

Table I. Complete comparisons among state-of-the-art driving world models in terms of *Subject Fidelity* in WorldLens.

$\mathcal{S}_{\mathrm{SF}}(\cdot)$	**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 (↑)	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 I. Examples of "good" and "bad" generation qualities in terms of *Subject Fidelity* in WorldLens.

A.2. Subject Coherence

A.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.

A.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:

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

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.

A.2.3. Implementation Details

We compute *Subject Coherence* using embeddings extracted from the Cross-Video ReID model of Zuo et al. [36]. 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.

A.2.4. Examples

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

A.2.5. Evaluation & Analysis

Tab. II provides the complete results of models in terms of *Subject Coherence*.

Table II. Complete comparisons among 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 II. Examples of "good" and "bad" generation qualities in terms of *Subject Coherence* in WorldLens.

A.3. Subject Consistency

A.3.1. Definition

Subject Consistency measures the temporal stability of object-level semantics and structural details. It focuses on fine-grained appearance and geometric regularity through DINO features [3], evaluating whether dynamic subjects maintain consistent texture, shape, and spatial structure over time. High subject consistency indicates that the model preserves the semantic identity and visual integrity of objects throughout motion, avoiding flickering or deformation.

A.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:**

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

  which measures the average cosine similarity between consecutive frame embeddings.
- **Temporal Jitter Index (TJI):**

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

  which measures normalized second-order fluctuations (lower is smoother).
- **Motion-Rate Similarity (MRS):**

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

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

The overall Subject Consistency score integrates these terms:

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

A high $\mathcal{S}_{\text{SC}}$ score implies that object-level features evolve smoothly over time, maintaining consistent structural integrity.

A.3.3. Implementation Details

We extract frame-wise features using DINO ViT-B/16 [3]. These normalized embeddings are used to compute adjacent-frame similarity, temporal jitter, and motion alignment against the corresponding ground-truth videos.

A.3.4. Examples

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

A.3.5. Evaluation & Analysis

Tab. III provides the complete results of models in terms of *Subject Consistency*.

Table III. Complete comparisons among state-of-the-art driving world models in terms of *Subject Consistency* in WorldLens.

$\mathcal{S}_{\text{SC}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	$\mathcal{X}$**-Scene** [NeurIPS'25]	**Empirical Max**
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 III. Examples of "good" and "bad" generation qualities in terms of *Subject Consistency* in WorldLens.

A.4. Depth Discrepancy

A.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.

A.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:

$$\text{DD}(y_j) = \frac{1}{T-1} \sum_{t=1}^{T-1} \| f_j^{(t)} - f_j^{(t+1)} \|_2 \, .$$

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

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

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

A.4.3. Implementation Details

Depth maps for both generated and ground-truth videos are obtained using Video DepthAnything [4]. The predicted depths are directly used to compute the per-frame depth discrepancy.

A.4.4. Examples

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

A.4.5. Evaluation & Analysis

Tab. IV provides the complete results of models in terms of *Depth Discrepancy*.

Table IV. Complete comparisons among 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

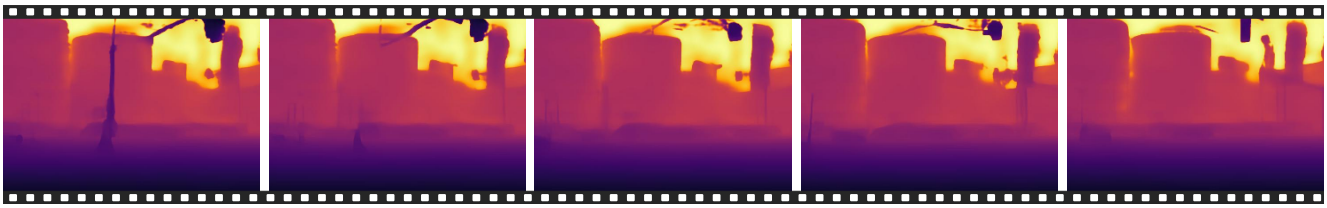
(a) Good example in the *Depth Discrepancy* dimension (Score: 4.43)

(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 IV. Examples of "good" and "bad" generation qualities in terms of *Depth Discrepancy* in WorldLens.

A.5. Temporal Consistency

A.5.1. Definition

Temporal Consistency quantifies the frame-to-frame stability of generated videos in a learned appearance space. Using a frozen CLIP encoder $\phi_{\text{CLIP}}(\cdot)$ [19], this metric captures whether visual representations evolve smoothly over time without abrupt changes or flickering. It measures three complementary aspects: (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.

A.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* (A.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:

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

The dataset-level metric averages per-video scores:

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

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.

A.5.3. Implementation Details

Temporal Consistency is evaluated using frame-wise features from CLIP ViT-B/32 [19] 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.

A.5.4. Examples

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

A.5.5. Evaluation & Analysis

Tab. V provides the complete results of models in terms of *Temporal Consistency*.

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

$\mathcal{S}_{\text{TC}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
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 V. Examples of "good" and "bad" generation qualities in terms of *Temporal Consistency* in WorldLens.

A.6. Semantic Consistency

A.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 $\psi_{\mathrm{SEG}}(\cdot)$, this metric quantifies how consistently pixel-wise labels, region structures, and global class distributions are preserved between consecutive frames. High consistency implies temporally coherent scene semantics without class flicker or unstable object boundaries.

A.6.2. Formulation

For each generated video $y_j = \{y_j^{(t)}\}_{t=1}^{T}$, we obtain frame-wise segmentation masks: $M_j^{(t)} = \psi_{\mathrm{SEG}}(y_j^{(t)}) \in \{0,\dots,C-1\}^{H\times W}$. 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:

$$\mathcal{S}_{\mathrm{LFR}}(y_j) = 1 - \tfrac{1}{T-1}\sum_{t=1}^{T-1} \frac{\sum_c \sum_{\mathbf{p}\in\Omega_c^{(t)}} \mathbf{1}[M_j^{(t+1)}(\mathbf{p})\neq c]}{\sum_c |\Omega_c^{(t)}|} .$$

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:

$$\mathcal{S}_{\mathrm{SAC}}(y_j) = \tfrac{1}{T-1}\sum_{t=1}^{T-1} \frac{\sum_c \sum_{(R,R')\in\pi_c^{(t)}} |R|\cdot \mathrm{IoU}(R,R')}{\sum_c \sum_{R\in\mathcal{R}_c^{(t)}} |R|} ,$$

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:

$$\mathcal{S}_{\mathrm{CDS}}(y_j) = 1 - \tfrac{1}{T-1}\sum_{t=1}^{T-1} \mathrm{JSD}\Big(p^{(t)} \,\|\, p^{(t+1)}\Big) .$$

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

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

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

A.6.3. Implementation Details

We obtain frame-wise semantic maps using the panoptic segmentation model from OpenSeeD [33]. The predicted segments are then converted to label masks via a fixed color palette and used to compute the temporal semantic consistency score.

A.6.4. Examples

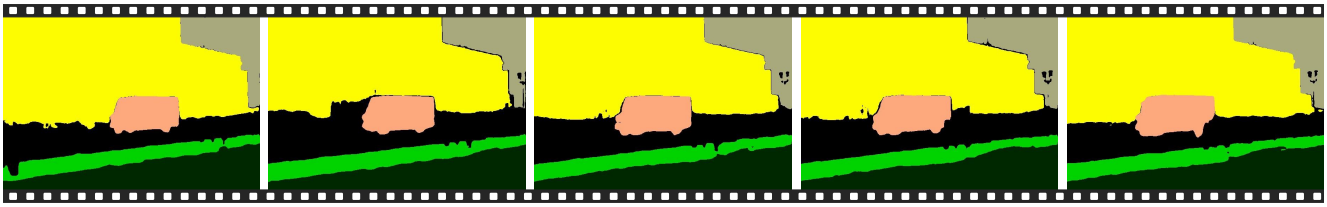
Fig. VI provides typical examples of videos with good and bad quality in terms of *Semantic Consistency*.

A.6.5. Evaluation & Analysis

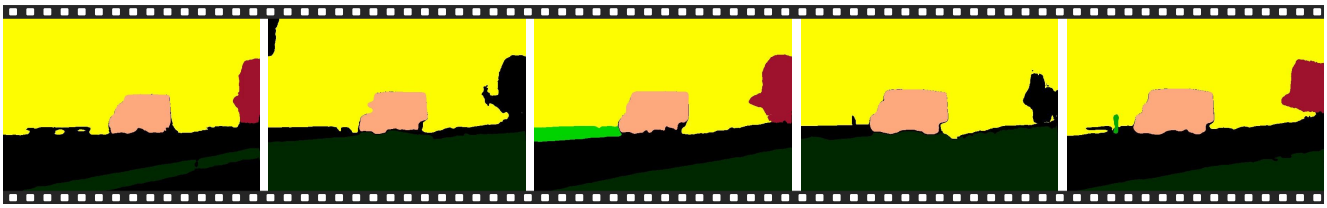
Tab. VI provides the complete results of models in terms of *Semantic Consistency*.

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

$\mathcal{S}_{\mathrm{SemC}}(\cdot)$	**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%



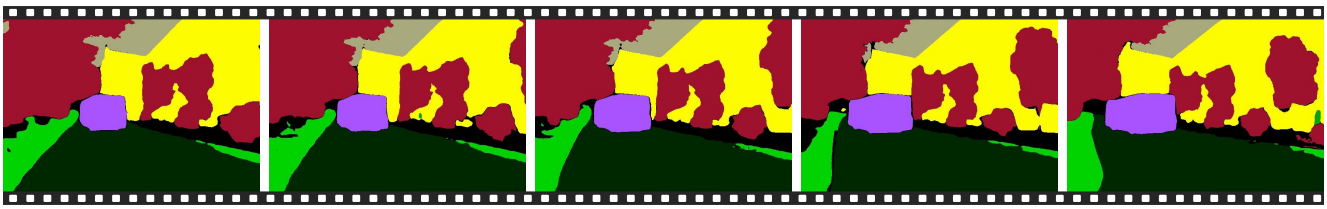
(a) Good example in the *Semantic Consistency* dimension (Score: 94.99%)



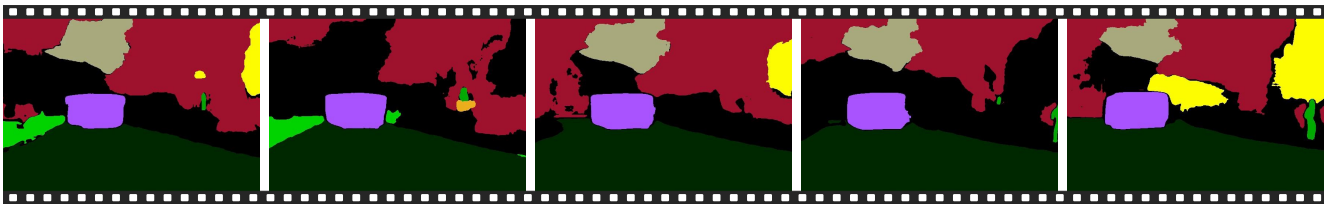
(b) Bad example in the *Semantic Consistency* dimension (Score: 74.70%)

(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 VI. Examples of "good" and "bad" generation qualities in terms of *Semantic Consistency* in WorldLens.

A.7. Perceptual Discrepancy

A.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 [24] trained on Kinetics [13].

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.

A.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}:

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

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:

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

A lower $\mathcal{S}_{\mathrm{PD}}$ 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.

A.7.3. Implementation Details

Perceptual Discrepancy is measured using Fréchet Video Distance (FVD). We extract video features with a pretrained I3D model [24] (Kinetics-400 [13]) following the VideoGPT [30] protocol, and compute FVD between ground-truth and generated feature distributions.

A.7.4. Examples

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

A.7.5. Evaluation & Analysis

Tab. VII provides the complete results of models in terms of *Perceptual Discrepancy*.

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

$\mathcal{S}_{\mathrm{PD}}(\cdot)$	**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 VII. Examples of "good" and "bad" generation qualities in terms of *Perceptual Discrepancy* in WorldLens.

A.8. Cross-View Consistency

A.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.

A.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:

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

Higher $\mathcal{S}_{\mathrm{CVC}}$ 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.

A.8.3. Implementation Details

The *Cross-View Consistency* score is computed by extracting frame-wise sparse correspondences using the pretrained LoFTR local feature matcher [23]. Matched keypoints across views are used to assess geometric alignment between generated and ground-truth videos.

A.8.4. Examples

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

A.8.5. Evaluation & Analysis

Tab. VIII provides the complete results of models in terms of *Cross-View Consistency*.

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

$\mathcal{S}_{\mathrm{CVC}}(\cdot)$	**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 VIII. Examples of "good" and "bad" generation qualities in terms of *Cross-View Consistency* in WorldLens.

B. Aspect 2: Reconstruction

This aspect assesses the **reconstructability** of generated videos, how accurately a consistent 4D scene can be recovered from synthesized frames. 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. To this end, we employ a differentiable 4D reconstruction algorithm that optimizes scene geometry and radiance from the generated sequences, then re-renders the reconstructed model under both original and unseen camera poses.

B.1. Photometric Discrepancy

B.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 4D reconstruction pipeline based on recent neural rendering frameworks such as 4D Gaussian Splatting or NeRF-based [17] 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 [26] and LPIPS [34]. High photometric accuracy indicates that the generated frames exhibit temporally consistent appearance, lighting, and reflectance properties, supporting reliable inverse reconstruction into a coherent 4D radiance representation.

B.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: $\hat{y}_j^{(t)} = \mathrm{Render}\Big(\hat{\mathcal{R}}_j, \mathrm{Pose}(t)\Big)$, 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:

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

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

B.1.3. Implementation Details

We follow the OmniRe [5] preprocessing pipeline and default configuration on nuScenes [2], 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 [5] 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.

B.1.4. Examples

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

B.1.5. Evaluation & Analysis

Tab. IX provides the complete results of models in terms of *Photometric Discrepancy*.

Table IX. Complete comparisons among 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 IX. Examples of "good" and "bad" reconstruction qualities in terms of *Photometric Discrepancy* in WorldLens.

B.2. Geometric Discrepancy

B.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 (AbsRel) computed on regions defined by Grounded-SAM 2 [20] masks that isolate road surfaces and foreground objects. Lower values indicate that the reconstructed geometry from the generated video closely matches the ground-truth scene.

B.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:

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

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

B.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 [32]) as OmniRe [5], 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 [20], 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.

B.2.4. Example

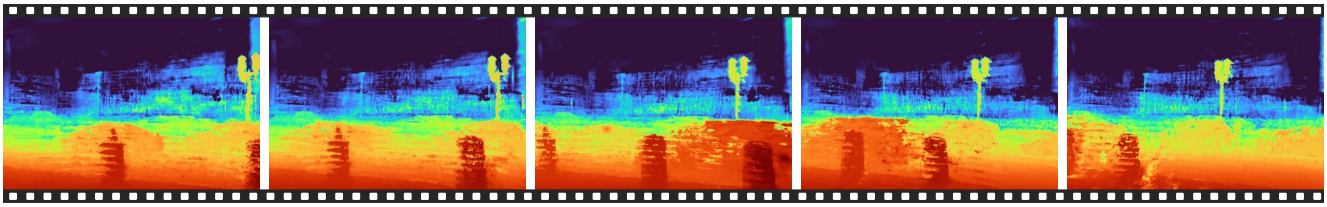
Fig. X provides typical examples of videos with good and bad quality in terms of *Geometric Discrepancy*.

B.2.5. Evaluation & Analysis

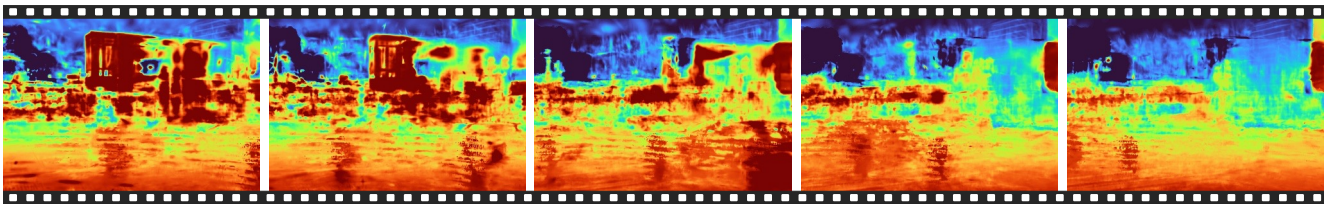
Tab. X provides the complete results of models in terms of *Geometric Discrepancy*.

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

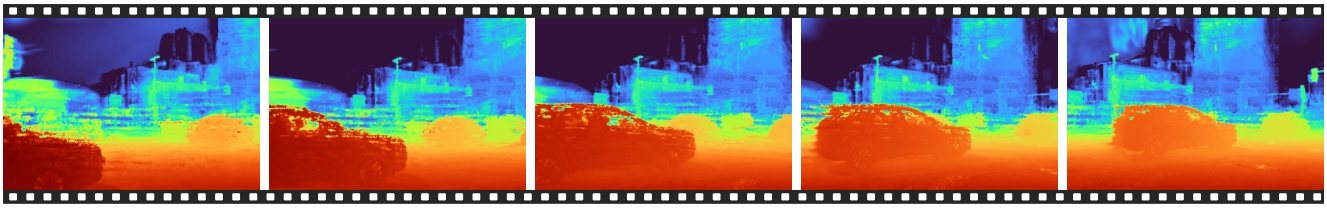
$\mathcal{S}_{\mathrm{GeoA}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
RMSE (↓)	4.116	4.166	2.869	3.130	2.969	3.594	-
Abs Rel (↓)	0.115	0.105	0.073	0.088	0.080	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	-



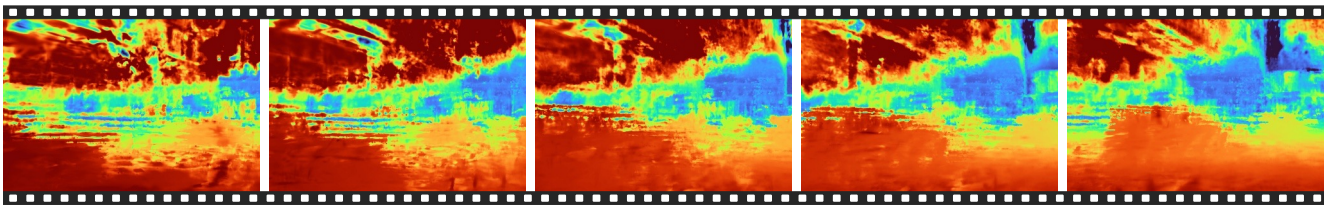
(a) Good example in the *Geometric Discrepancy* dimension (Score: 0.033)



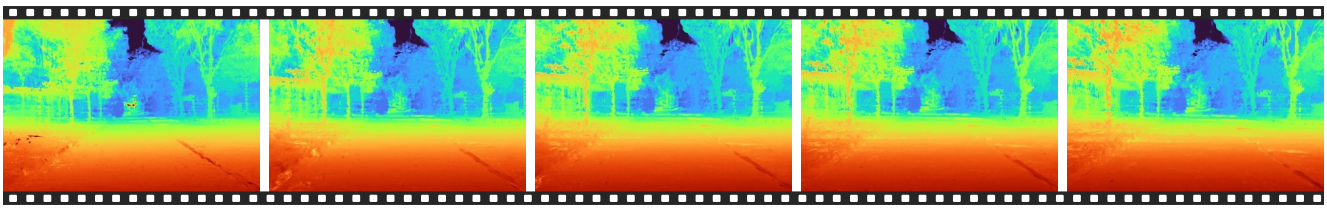
(b) Bad example in the *Geometric Discrepancy* dimension (Score: 0.156)



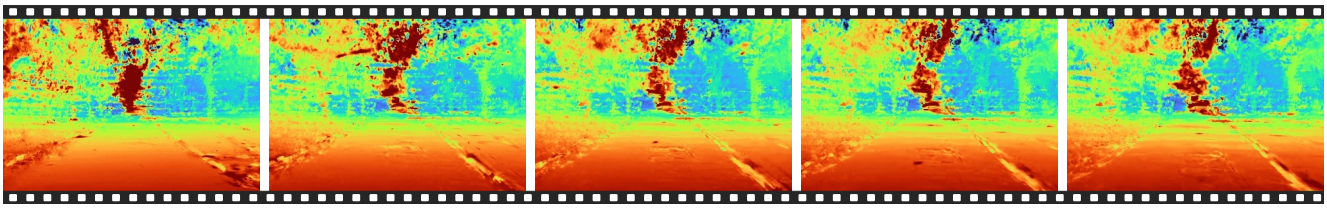
(c) Good example in the *Geometric Discrepancy* dimension (Score: 0.027)



(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 X. Examples of "good" and "bad" reconstruction qualities in terms of *Geometric Discrepancy* in WorldLens.

B.3. Novel-View Quality

B.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 [14]. A higher NVQ indicates that the rendered views exhibit sharper, more natural, and artifact-free visual details from unseen viewpoints.

Three novel-view trajectories are considered: `front_center_interp` 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` 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` generates a parallel-view sequence by shifting each front-camera pose (ID 0) laterally by a fixed offset distance 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.

B.3.2. Formulation

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

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

$$\boxed{\mathcal{S}_{\mathrm{NVQ}}(\mathcal{Y}) = \frac{1}{N_g T} \sum_{j=1}^{N_g} \sum_{t=1}^{T} q_j^{(t)}} \tag{11}$$

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

B.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 B.3.1, where the lateral offset is set to 1 m. Novel-view image quality is assessed using the pretrained MUSIQ model [14]. 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.

B.3.4. Examples

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

B.3.5. Evaluation & Analysis

Tab. XI provides the complete results of models in terms of *Novel-View Quality*.

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

$\mathcal{S}_{\mathrm{NVQ}}(\cdot)$	**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 XI. Examples of "good" and "bad" reconstruction qualities in terms of *Novel-View Quality* in WorldLens.

B.4. Novel-View Discrepancy

B.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.

B.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. Equation 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.

B.4.3. Implementation Details

The selection of novel viewpoints and the rendering configurations are kept consistent with those in Section B.3. For *Novel-View Discrepancy*, the calculation process follows the same setting of Section A.7.

B.4.4. Example

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

B.4.5. Evaluation & Analysis

Tab. XII provides the complete results of models in terms of *Novel-View Discrepancy*.

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

$\mathcal{S}_{\text{NVD}}(\cdot)$	**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 (↓)	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 XII. Examples of "good" and "bad" reconstruction qualities in terms of *Novel-View Discrepancy* in WorldLens.

C. 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.

C.1. Displacement Error

C.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.

C.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.

C.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 [12], 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.

C.1.4. Examples

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

C.1.5. Evaluation & Analysis

Tab. XIII provides the complete results of models in terms of *Displacement Error*.

Table XIII. Complete comparisons among 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 XIII. Examples of "good" and "bad" action-following performances in terms of *Displacement Error* in WorldLens.

C.2. Open-Loop Adherence

C.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 visual input in a non-reactive simulation environment. Following NAVSIM [7], 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.

The PDMS aggregates multiple sub-scores related to safety, progress, and comfort within a short open-loop simulation horizon, providing a more realistic proxy to closed-loop evaluation than traditional displacement metrics.

C.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-vehicle 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:

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

with default weights $\text{weight}_{\text{EP}}{=}\text{weight}_{\text{TTC}}{=}5$ and $\text{weight}_{\text{C}}{=}2$ as in [7]. We report the dataset-level score as the mean PDMS across all evaluated videos:

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

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

C.2.3. Implementation Details

We support two map environments, *singapore-onenorth* and *boston-seaport*, aligned with the DriveArena platform [31]. 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 [28] 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.

C.2.4. Examples

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

C.2.5. Evaluation & Analysis

Tab. XIV provides the complete results of models in terms of *Open-Loop Adherence*.

Table XIV. Complete comparisons among 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 XIV. Examples of "good" and "bad" action-following performances in terms of *Open-Loop Adherence* in WorldLens.

C.3. Route Completion

C.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.

C.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 [7, 31], *Route Completion* is defined as the ratio between the completed and total distances:

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

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

$$\boxed{\mathcal{S}_{\text{RC}}(\mathcal{Y}) = \frac{1}{N_g} \sum_{j=1}^{N_g} \text{RC}(y_j)} \tag{17}$$

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

C.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.

C.3.4. Examples

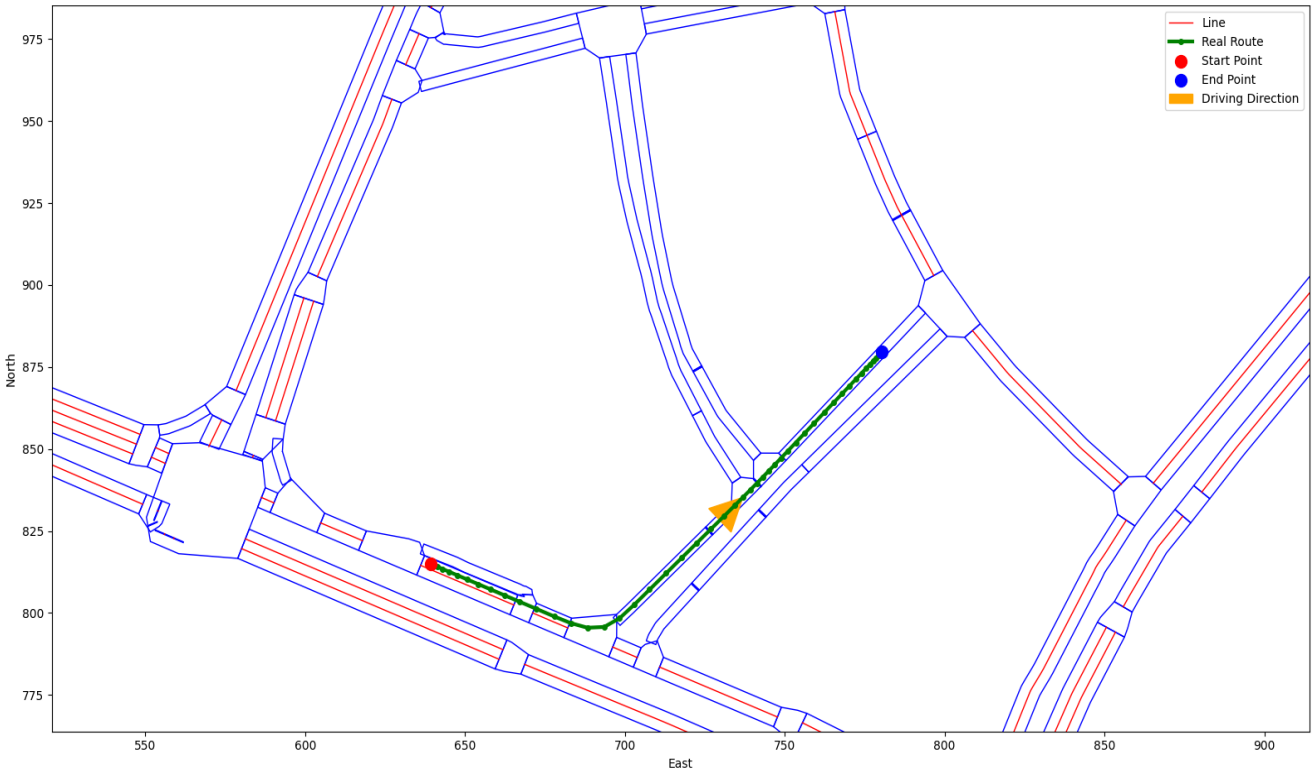
Fig. XV provides typical examples of videos with good and bad quality in terms of *Route Completion*.

C.3.5. Evaluation & Analysis

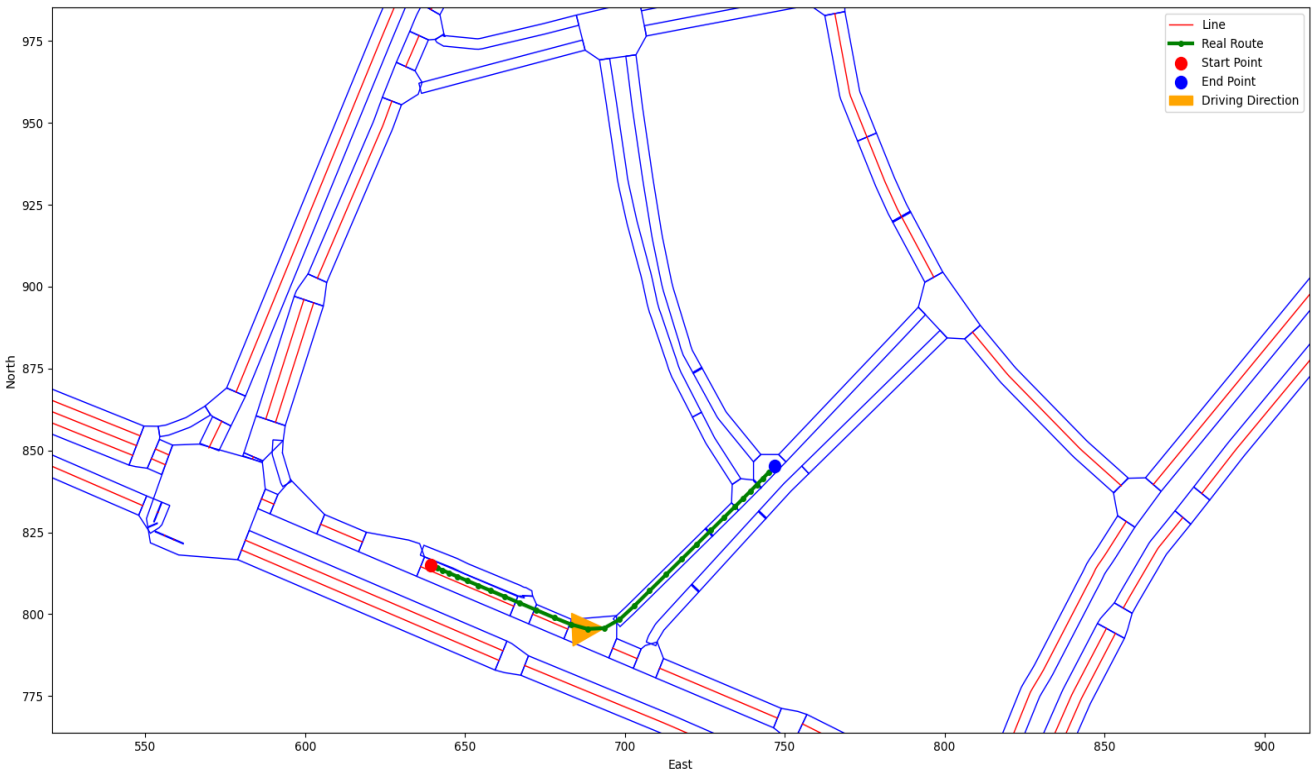
Tab. XV provides the complete results of models in terms of *Route Completion*.

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

$\mathcal{S}_{\text{RC}}(\cdot)$	**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 XV. Examples of "good" and "bad" action-following performances in terms of *Route Completion* in WorldLens.

C.4. Closed-Loop Adherence

C.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)* [31], 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.

C.4.2. Formulation

Given the PDMS and RC metrics defined in Equation C.2.2 and Equation C.3.2, the Arena Driving Score is computed as follows:

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

where $\mathrm{RC} \in [0, 1]$ 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:

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

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

C.4.3. Implementation Details

Closed-Loop Adherence shares the same experiment environment with *Route Completion*. The implementation details can be found in Section C.3.

C.4.4. Examples

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

C.4.5. Evaluation & Analysis

Tab. XVI provides the complete results of models in terms of *Closed-Loop Adherence*.

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

$\mathcal{S}_{\mathrm{ADS}}(\cdot)$	**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 XVI. Examples of "good" and "bad" action-following performances in terms of *Closed-Loop Adherence* in WorldLens.

D. Aspect 4: Downstream Task

In this section, we evaluate the **downstream task utility** of generated driving 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 domain reflects the distribution gap introduced by generation. Thus, the closer the downstream performance on generated data is to that on real data, the higher the overall quality, realism, and utility of the generative model.

D.1. Map Segmentation

D.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 $\psi_{\text{BEV}}(\cdot)$ 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.

D.1.2. Formulation

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

$$S_{\text{BEV}}^{(t)}(y_j) = \frac{1}{C_{\text{BEV}}} \sum\nolimits_{c=1}^{C_{\text{BEV}}} \frac{|\hat{B}_j^{(t,c)} \cap B_j^{\text{gt}(t,c)}|}{|\hat{B}_j^{(t,c)} \cup B_j^{\text{gt}(t,c)}|}.$$

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

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

where C_{BEV} is the number of BEV categories and (H_b, W_b) the BEV map resolution. Higher $\mathcal{S}_{\text{Seg}}$ indicates that generated frames enable accurate BEV map prediction consistent with real scenes.

D.1.3. Implementation Details

For *Map Segmentation*, we employ the pretrained BEVFusion multi-task model of Liu et al. [16], using its camera-only configuration with a ResNet-101 [11] 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.

D.1.4. Examples

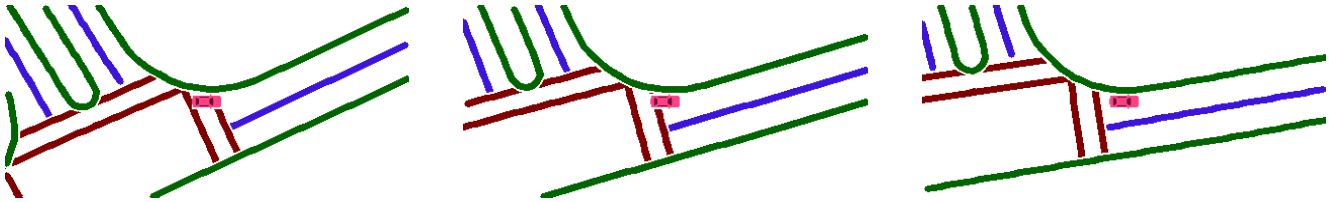
Fig. XVII provides typical examples of videos with good and bad quality in terms of *Map Segmentation*.

D.1.5. Evaluation & Analysis

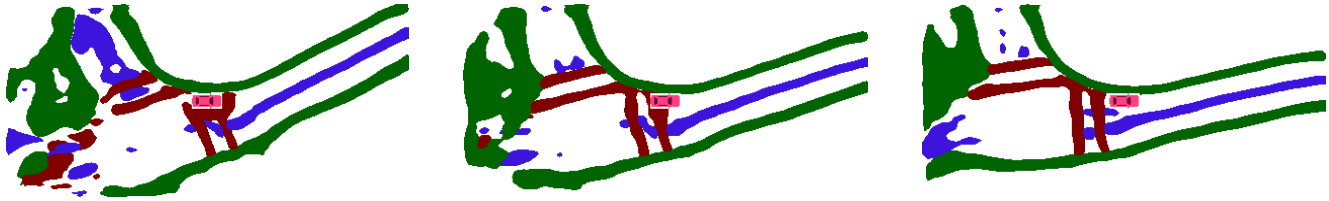
Tab. XVII provides the complete results of models in terms of *Map Segmentation*.

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

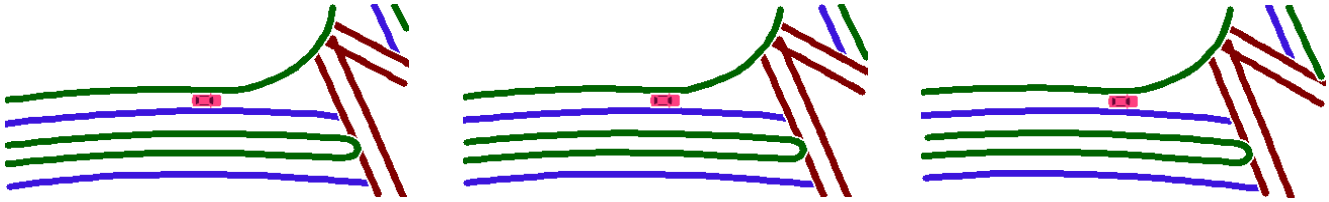
$\mathcal{S}_{\text{Seg}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	$\mathcal{X}$**-Scene** [NeurIPS'25]	**Empirical Max**
Divider (↑)	23.39%	34.44%	39.69%	31.88%	41.26%	31.84%	46.08%
Ped. Crossing (↑)	9.77%	21.18%	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%



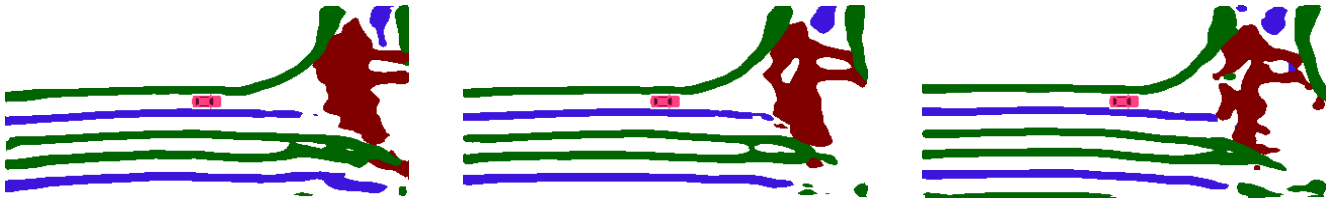
(a) Good example in the *Map Segmentation* dimension (Score: 100.00%)



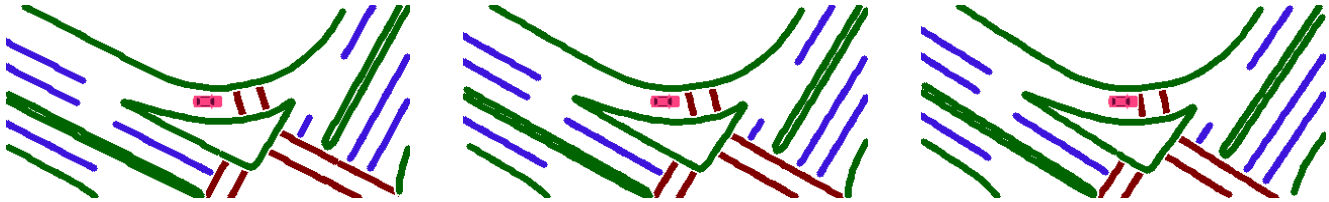
(b) Bad example in the *Map Segmentation* dimension (Score: 9.46%)



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

D.2. 3D Object Detection

D.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 [2], 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.

D.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 [15, 16] using standard matching and error terms. The dataset-level 3D detection score averages these values across all generated frames:

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

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

D.2.3. Implementation Details

The 3D detection evaluation uses the same pretrained BEVFusion model as in Section D.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.

D.2.4. Examples

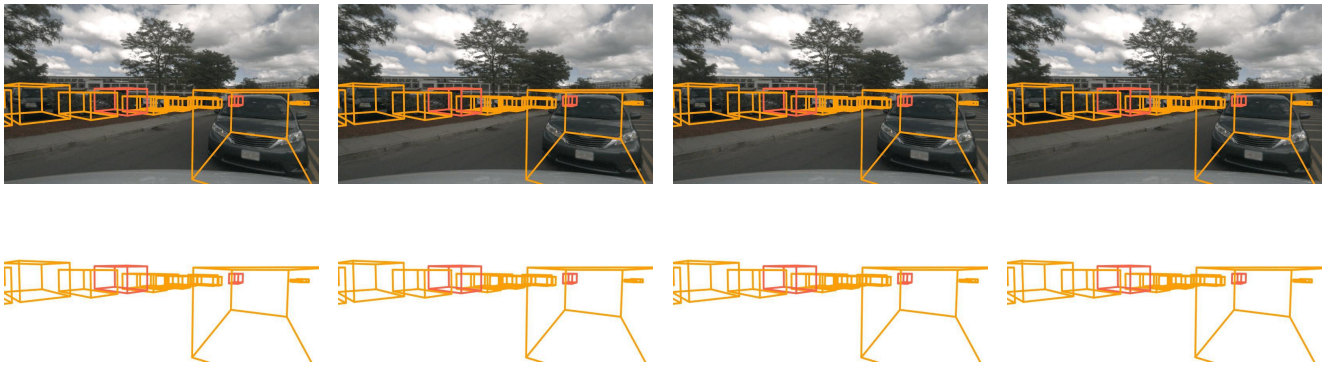
Fig. XVIII provides typical examples of videos with good and bad quality in terms of *3D Object Detection*.

D.2.5. Evaluation & Analysis

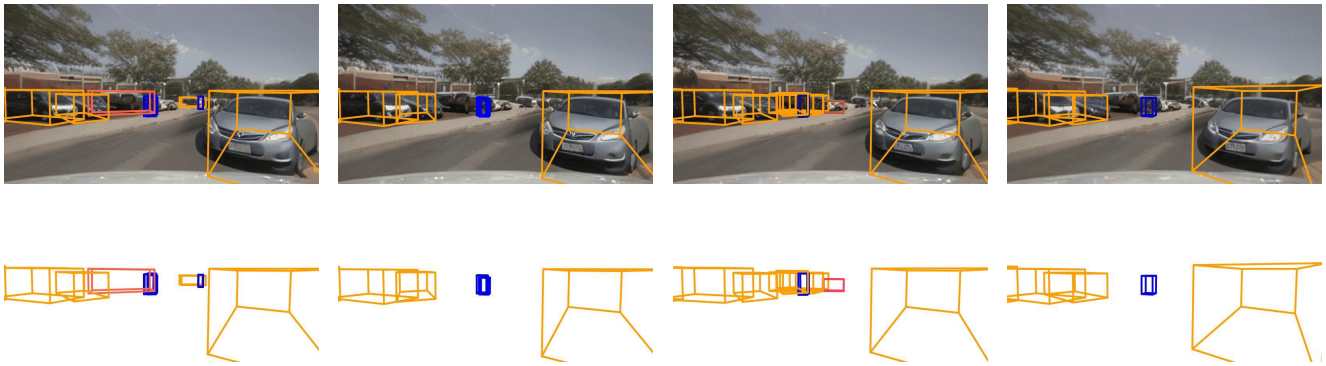
Tab. XVIII provides the complete results of models in terms of *3D Object Detection*.

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

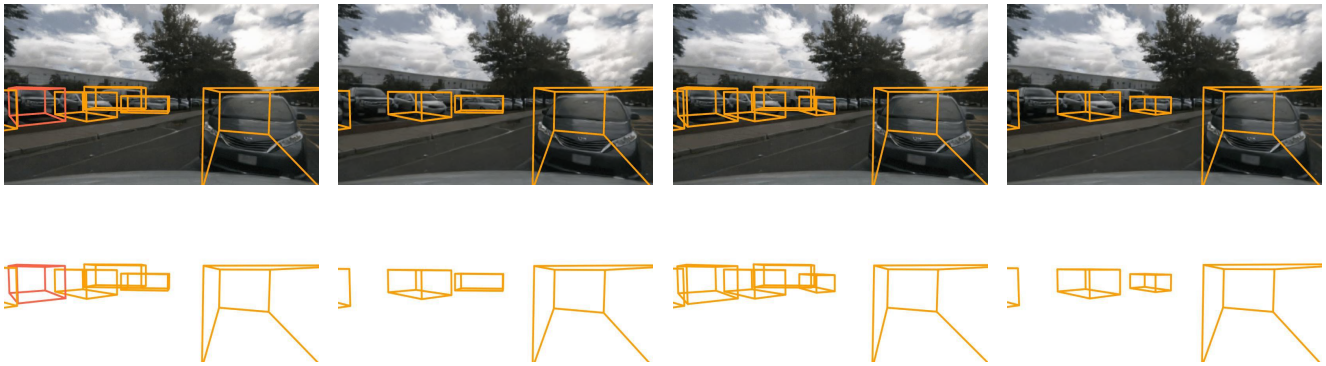
$\mathcal{S}_{\mathrm{Det}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
mAP (↑)	0.1178	0.1636	0.1961	0.944	0.2242	0.1562	0.3657
mATE (↓)	0.9435	0.9469	0.8443	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



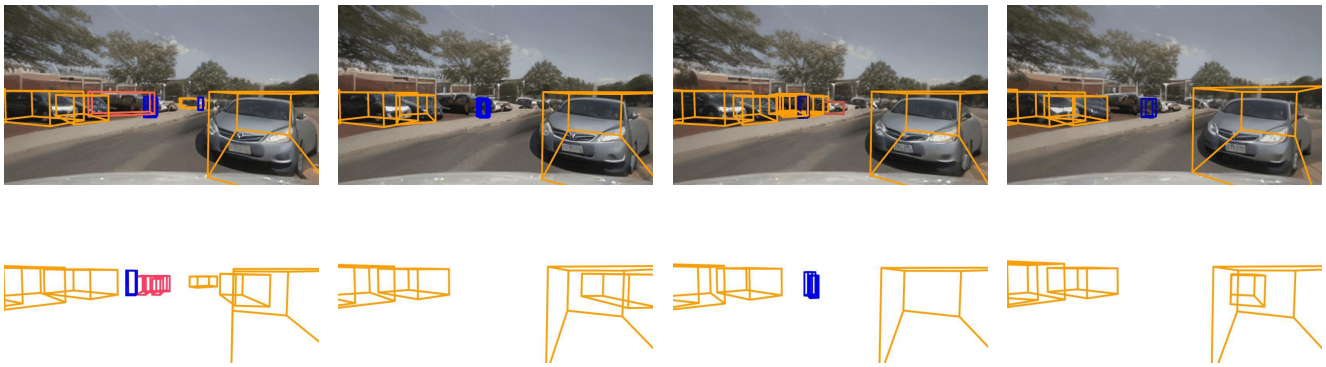
(a) Good example in the *3D Object Detection* dimension



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

D.3. 3D Object Tracking

D.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 $\psi_{\mathrm{TRK}}(\cdot)$, trained on ground-truth sequences, is applied to each generated video to estimate 3D trajectories of dynamic objects.

Following the nuScenes tracking protocol [2], 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.

D.3.2. Formulation

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

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

and $\mathcal{T}_j^{\mathrm{gt}}$ denotes the corresponding ground-truth trajectories. Tracking accuracy is evaluated using the official nuScenes metrics [2], 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:

$$\boxed{\mathcal{S}_{\mathrm{Trk}}(\mathcal{Y}) = \frac{1}{N_g}\sum_{j=1}^{N_g}\mathrm{AMOTA}(y_j)} \tag{21}$$

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

D.3.3. Implementation Details

We evaluate the 3D object tracking performance using the pretrained camera-only ADA-Track [9], following its official nuScenes configuration. The tracker is run directly on the generated multi-view videos.

D.3.4. Examples

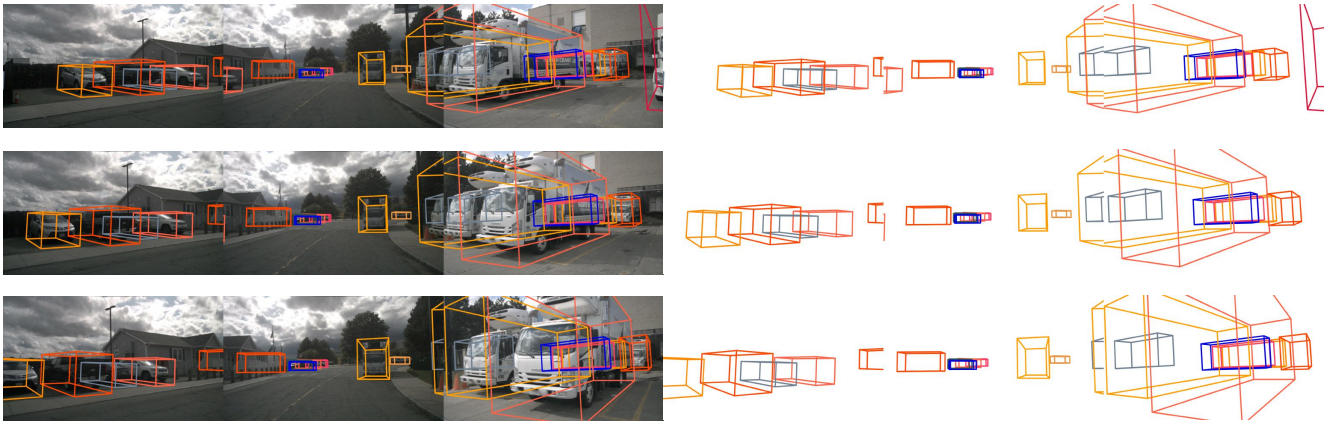
Fig. XIX provides typical examples of videos with good and bad quality in terms of *3D Object Tracking*.

D.3.5. Evaluation & Analysis

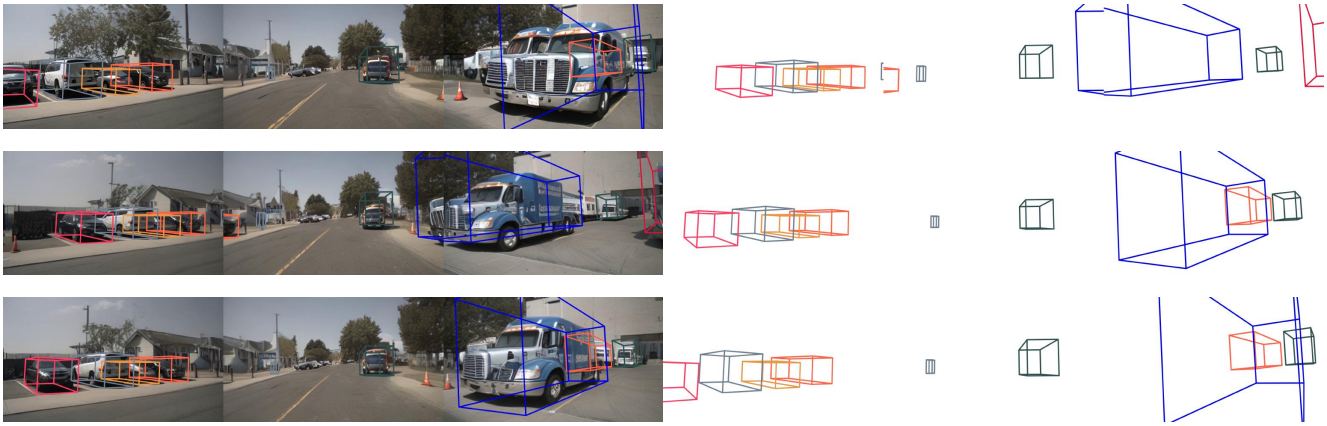
Tab. XIX provides the complete results of models in terms of *3D Object Tracking*.

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

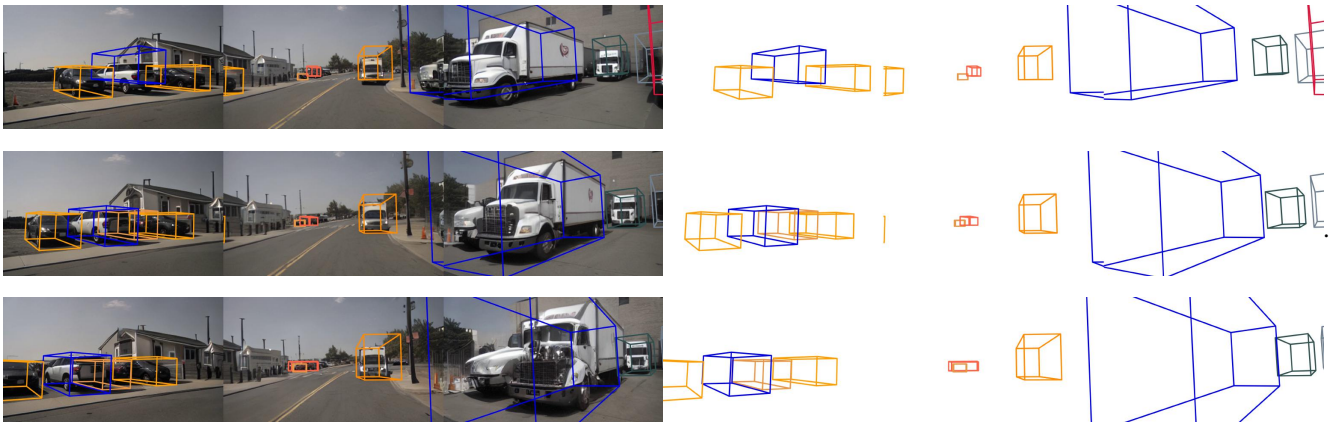
$\mathcal{S}_{\mathrm{Trk}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
AMOTP (↓)	1.843	1.812	1.778	1.848	1.705	1.829	1.405
Recall (↑)	15.50%	16.10%	19.50%	12.40%	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 XIX. Examples of "good" and "bad" downstream task performances in terms of *3D Object Tracking* in WorldLens.

D.4. Occupancy Prediction

D.4.1. Definition

Occupancy Prediction evaluates whether generated videos enable accurate 3D reconstruction of scene geometry and semantics. We adopt the RayIoU metric [25], 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. Higher RayIoU indicates more accurate and depth-consistent semantic occupancy.

D.4.2. Formulation

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

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

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 $(d_r^{\mathrm{gt}}, c_r^{\mathrm{gt}})$. A prediction is correct if $\hat{c}_r = c_r^{\mathrm{gt}}$ and $|\,\hat{d}_r - d_r^{\mathrm{gt}}\,| \le \delta$. The RayIoU at tolerance δ is defined as:

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

and the mean RayIoU (mRayIoU) aggregates multiple tolerances:

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

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

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

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

D.4.3. Implementation Details

We perform occupancy prediction using the pretrained SparseOcc model [25]. 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.

D.4.4. Examples

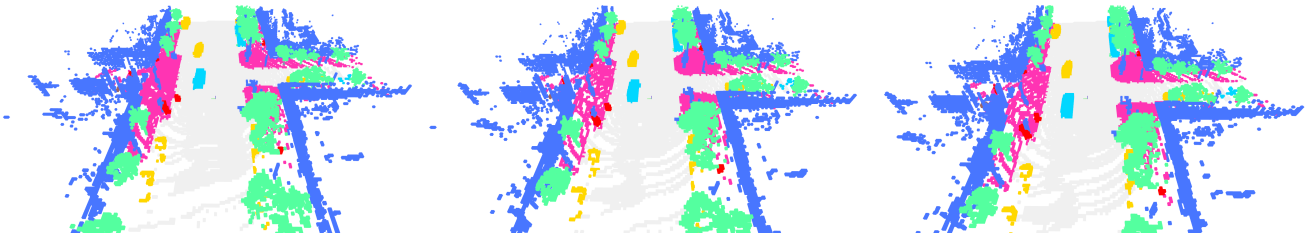
Fig. XX provides typical examples of videos with good and bad quality in terms of *Occupancy Prediction*.

D.4.5. Evaluation & Analysis

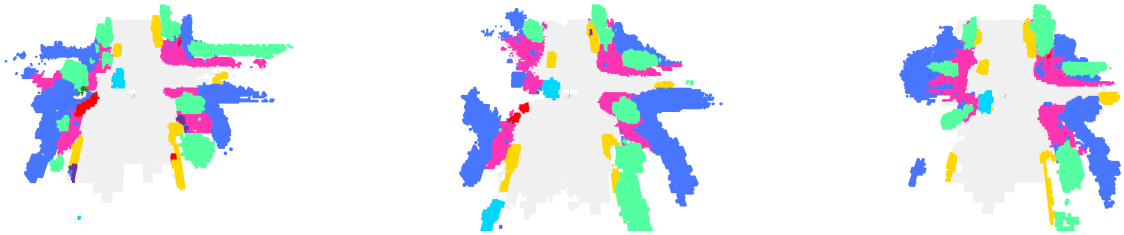
Tab. XX provides the complete results of models in terms of *Occupancy Prediction*.

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

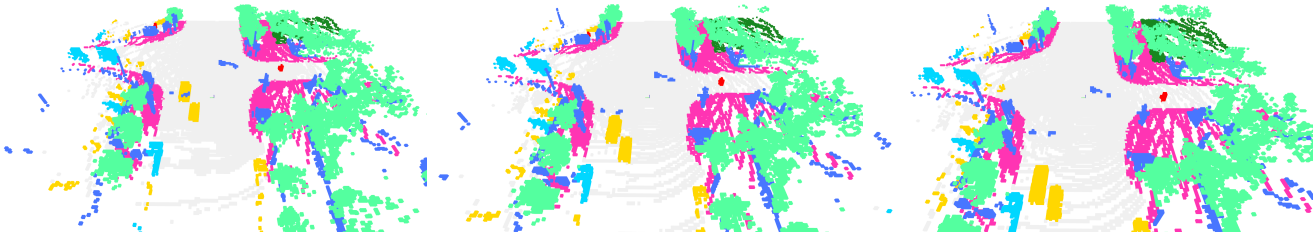
$\mathcal{S}_{\mathrm{Occ}}(\cdot)$	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
RayIoU@1m (↑)	17.24%	18.24%	20.32%	17.78%	18.81%	18.00%	29.04%
RayIoU@2m (↑)	23.53%	24.10%	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%



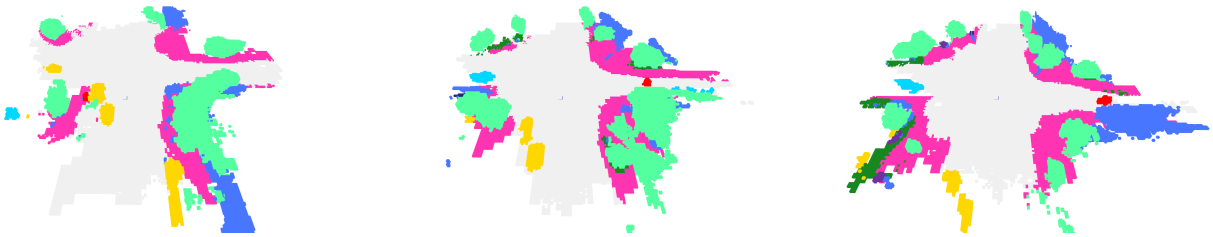
(a) Good example in the *Occupancy Prediction* dimension (Score: 100.00%)



(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 XX. Examples of "good" and "bad" downstream task performances in terms of *Occupancy Prediction* in WorldLens.

E. Aspect 5: Human Preference

In this section, we complement the automatic metrics with human-centered evaluation. 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.

E.1. World Realism - Overall Realism

Overall Realism measures the global visual believability of the entire scene. Annotators judge whether the generated video "looks like a real-world driving recording". Annotators are instructed to judge each clip according to 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 between 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.

E.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 composition; lighting and 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	-

E.1.2. Examples

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

E.1.3. Evaluation & Analysis

Tab. XXI provides the complete results of models in terms of *Overall Realism*.

Table XXI. Complete comparisons among 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 XXI. Examples of "good" and "bad" human preference alignments in terms of *Overall Realism* in WorldLens.

E.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.

E.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	-

E.2.2. Examples

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

E.2.3. Evaluation & Analysis

Tab. XXII provides the complete results of models in terms of *Vehicle Realism*.

Table XXII. Complete comparisons among 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 XXII. Examples of "good" and "bad" human preference alignments in terms of *Vehicle Realism* in WorldLens.

E.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.

E.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	-

E.3.2. Examples

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

E.3.3. Evaluation & Analysis

Tab. XXIII provides the complete results of models in terms of *Pedestrian Realism*.

Table XXIII. Complete comparisons among 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 XXIII. Examples of "good" and "bad" human preference alignments in terms of *Pedestrian Realism* in WorldLens.

E.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, cones) 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 large unexplained flashes, patches of incoherent reflection, or abrupt global brightness jumps.

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

E.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	-

E.4.2. Examples

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

E.4.3. Evaluation & Analysis

Tab. XXIV provides the complete results of models in terms of *Physical Plausibility*.

Table XXIV. Complete comparisons among 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 Plausibilit* dimension (Score: 10)

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

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

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

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

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

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

E.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 *3D & 4D Consistency* indicates that 3D box trajectories in generated videos closely track those in real scenes, both spatially and temporally.

E.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	-

E.5.2. Examples

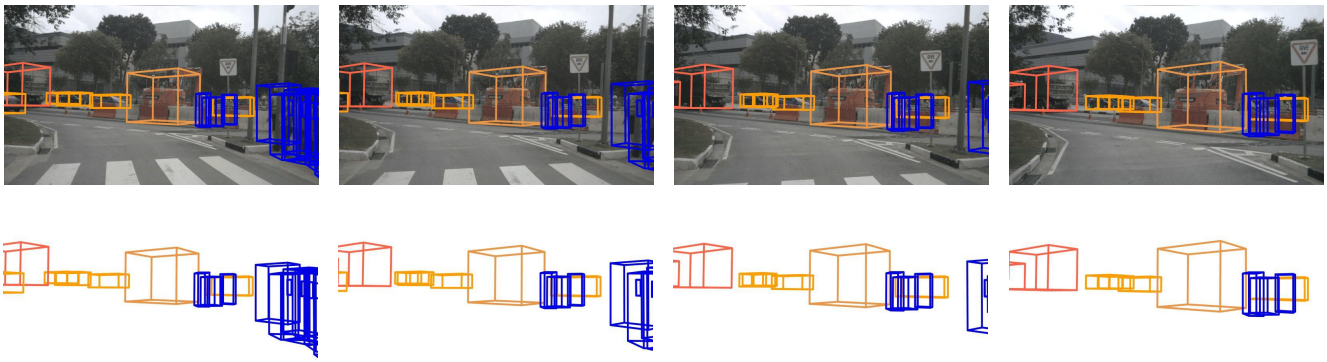
Fig. XXV provides typical examples of videos with good and bad quality in terms of *3D & 4D Consistency*.

E.5.3. Evaluation & Analysis

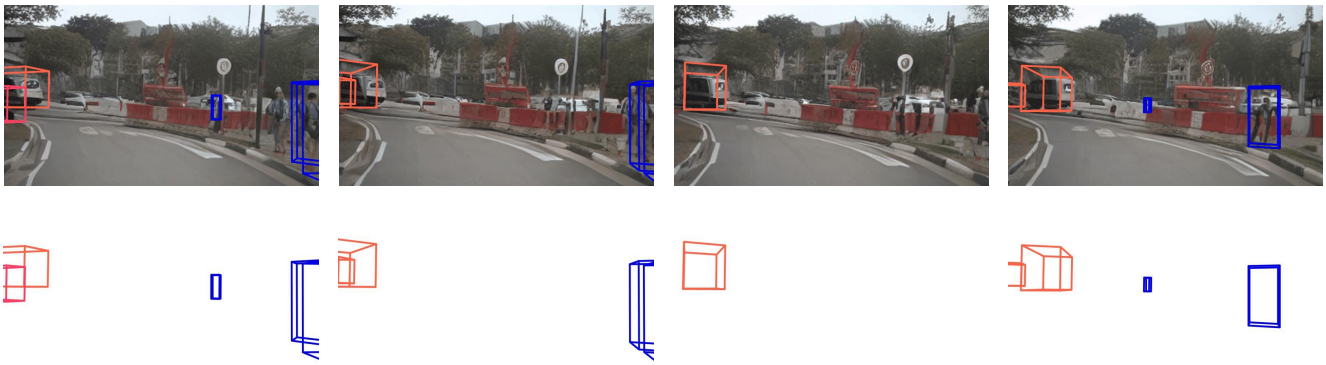
Tab. XXV provides the complete results of models in terms of *3D & 4D Consistency*.

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

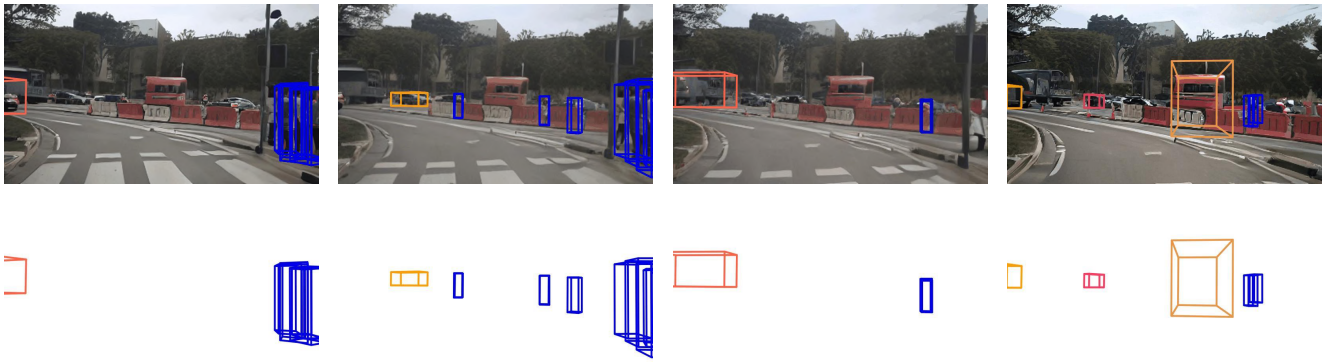
3D & 4D Consistency	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
min	2.0	2.0	2.0	2.0	2.0	2.0	-
max	10.0	10.0	10.0	10.0	10.0	10.0	-
mean	2.455	2.743	2.751	2.920	2.961	2.431	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	-



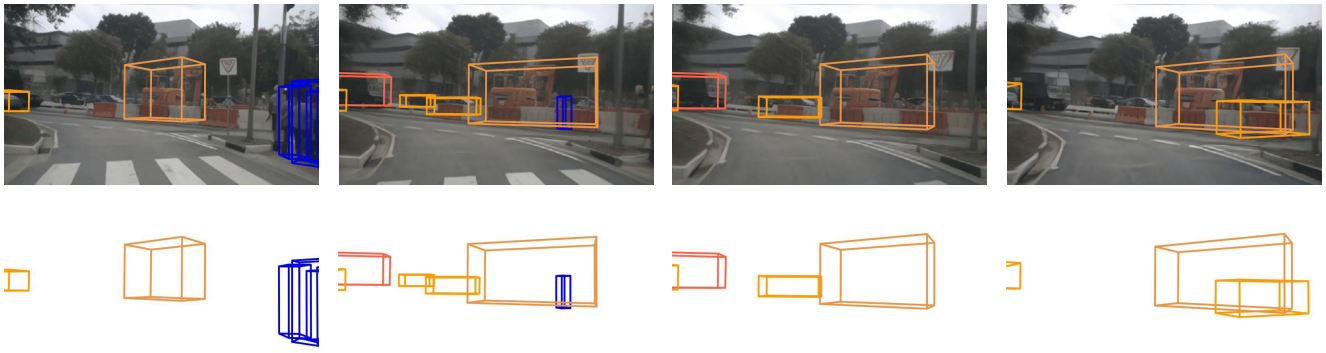
(a) Good example in the *3D & 4D Consistency* dimension (Score: 10)



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

E.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 such as 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 vulnerable 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.

E.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 but do not dominate the scene.
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	-

E.6.2. Examples

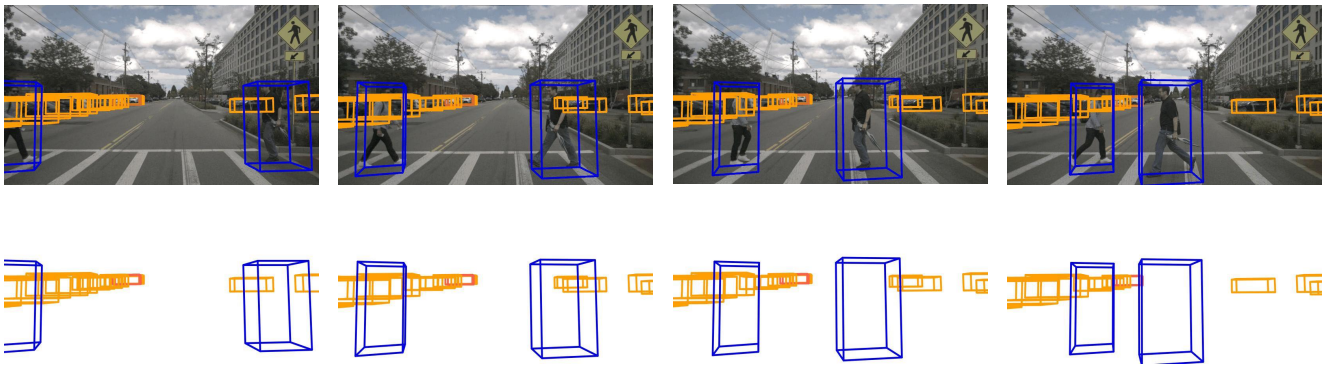
Fig. XXVI provides typical examples of videos with good and bad quality in terms of *Behavioral Safety*.

E.6.3. Evaluation & Analysis

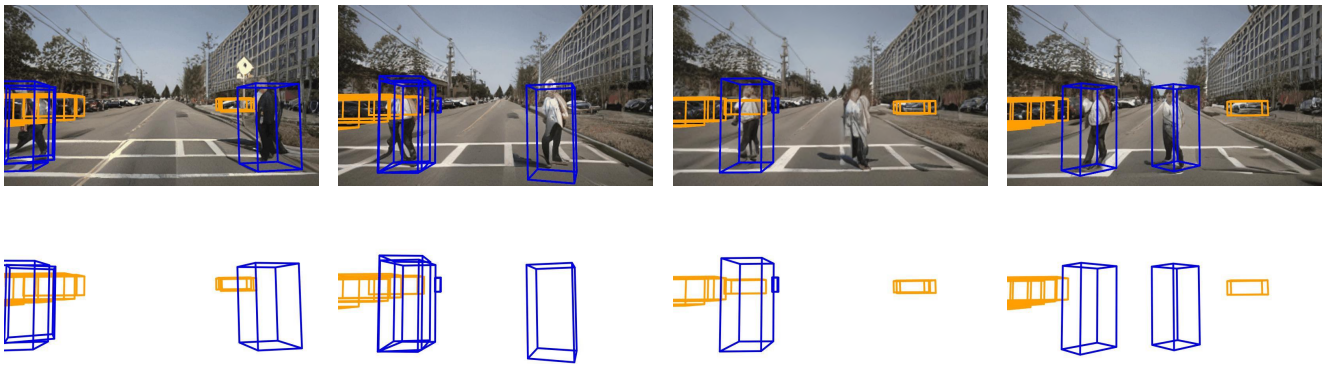
Tab. XXVI provides the complete results of models in terms of *Behavioral Safety*.

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

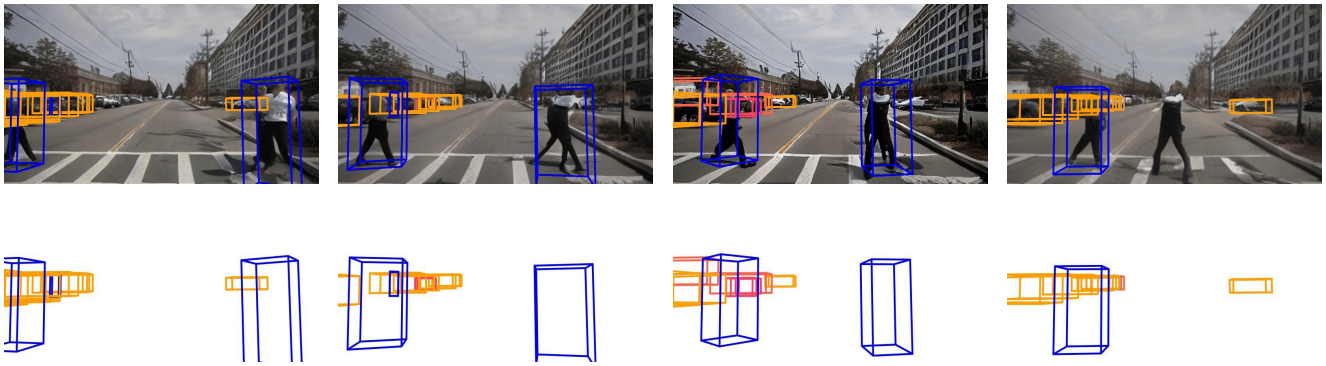
Behavioral Safety	**MagicDrive** [ICLR'24]	**DreamForge** [arXiv'24]	**DriveDreamer-2** [AAAI'25]	**OpenDWM** [CVPR'25]	**DiST-4D** [ICCV'25]	**$\mathcal{X}$-Scene** [NeurIPS'25]	**Empirical Max**
min	2.0	2.0	2.0	2.0	2.0	2.0	-
max	4.0	4.0	10.0	8.0	6.0	4.0	-
mean	2.306	2.290	2.533	2.598	2.591	2.318	10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



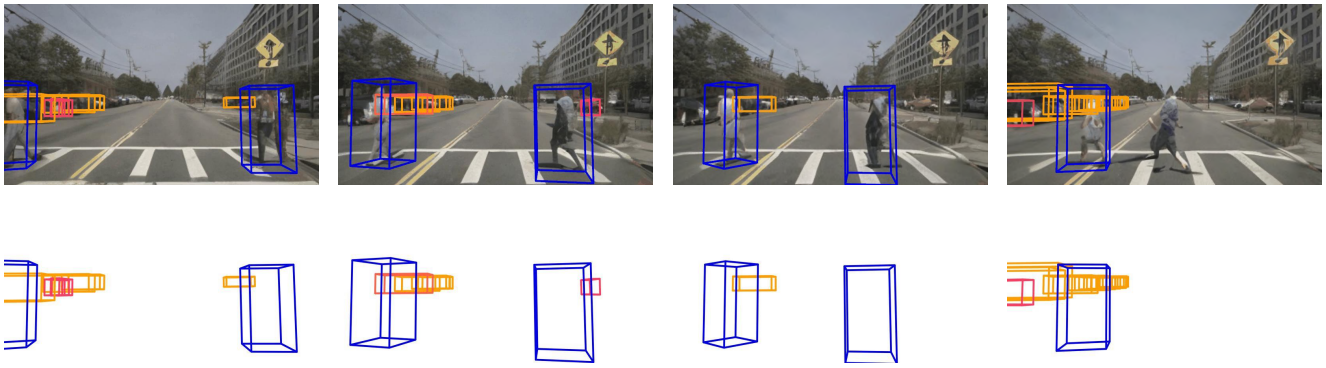
(a) Good example in the *Behavioral Safety* dimension (Score: 10)



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

F. 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.

F.1. Agent Architecture

The **WorldLens-Agent** is a vision-language critic built on Qwen3-VL-8B [1] and trained to evaluate generated videos along human-centered dimensions, including overall realism, 3D consistency, physical plausibility, and behavioral safety.

As shown in Fig. XXVII, 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.

F.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.

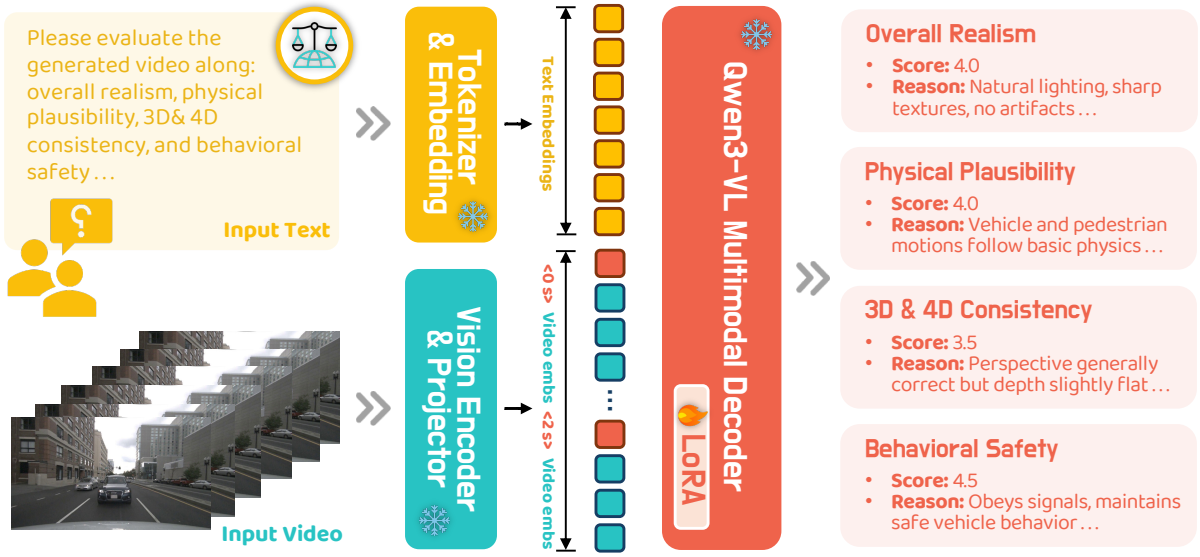


Figure XXVII. The architecture of the proposed **WorldLens-Agent** for auto-evaluation of generated driving videos.

F.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.

F.4. Consistency Alignment

Since original human-rated data are used for training, we additionally annotate a *held-out* set of 100 videos for agent evaluation. As shown in Table XXVII, WorldLens-Agent achieves a strong and consistent agreement with human scores across four held-out dimensions (H.1-H.4): Spearman ρ ranges from $\mathbf{0.81}$ to $\mathbf{0.91}$ and Kendall τ ranges from $\mathbf{0.72}$ to $\mathbf{0.84}$, with a macro-average of $\rho = \mathbf{0.86}$ and $\tau = \mathbf{0.76}$.

Beyond rank correlation, it also shows high absolute-score consistency (macro CCC $= \mathbf{0.80}$; MAE $= \mathbf{0.83}$; RMSE $= \mathbf{1.25}$), and high tolerance-level accuracy (Within $-$ 1 $= \mathbf{0.82}$, Within $-$ 2 $= \mathbf{0.93}$). These results validate that the agent's judgments are robust and not limited to qualitative alignment on a single external model.

F.5. Qualitative Assessment

Fig. XXVIII and Fig. XXIX present additional qualitative evaluations produced by the **WorldLens-Agent** on challenging driving scenarios, including **out-of-distribution** videos rendered or produced by Gen3C [22], Cosmos-Drive [21], and the CARLA [10] simulator. These examples illustrate the agent's ability to generalize 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 Fig. XXVIII, 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.

Fig. XXIX 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.

Table XXVII. Quantitative experiment on the correlation between WorldLens-Agent and human scores.

Dim	ρ	τ	CCC	MAE	RMSE	W1	W2
H.1	0.81	0.72	0.79	0.70	0.99	0.88	0.99
H.2	0.84	0.74	0.78	0.78	1.21	0.85	0.92
H.3	0.87	0.76	0.80	0.87	1.35	0.79	0.91
H.4	0.91	0.84	0.84	0.99	1.48	0.77	0.91
macro	0.86	0.76	0.80	0.83	1.25	0.82	0.93

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."
}
```

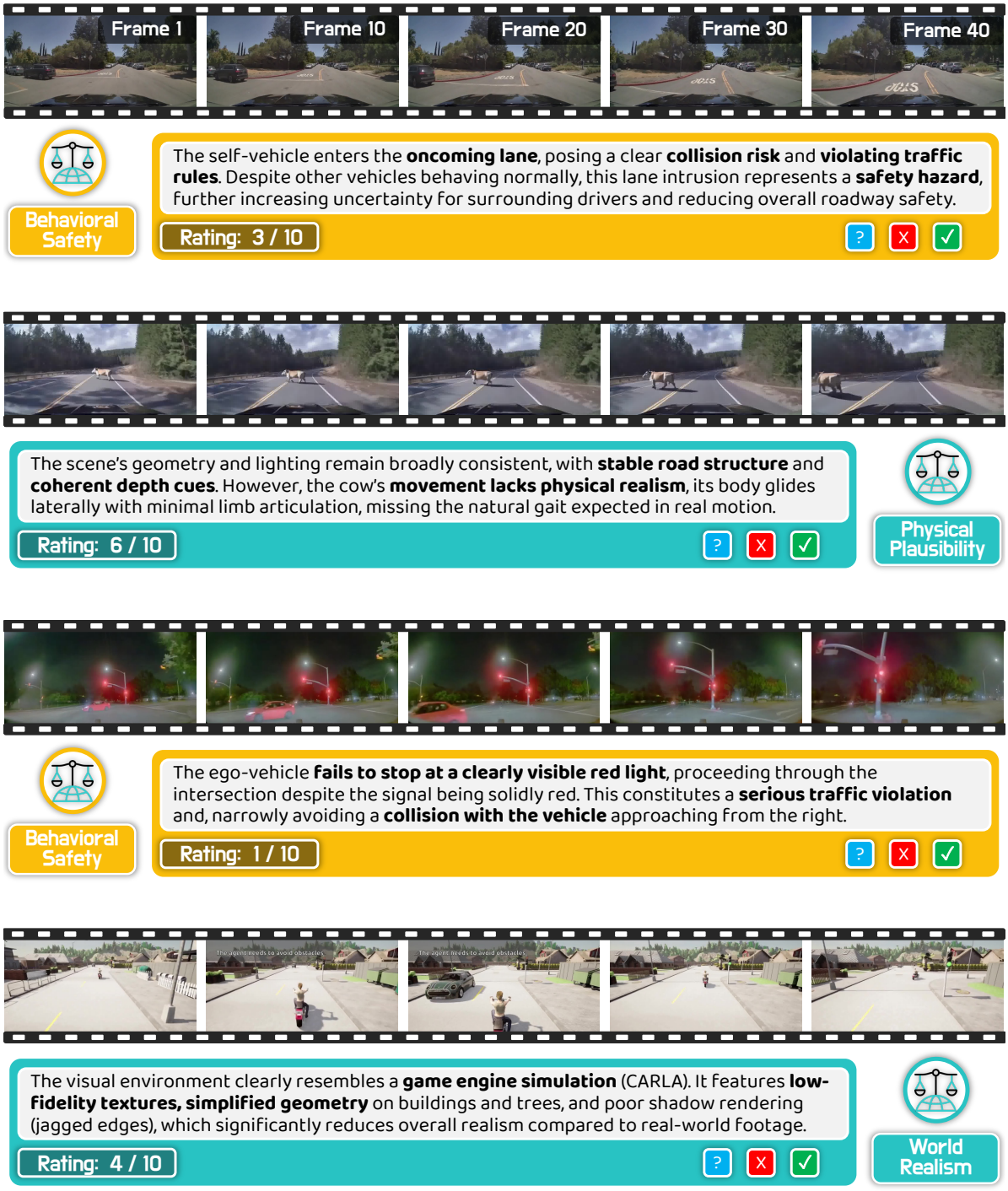


Figure XXVIII. Additional qualitative assessments of the **WorldLens-Agent** evaluation on challenging driving conditions.

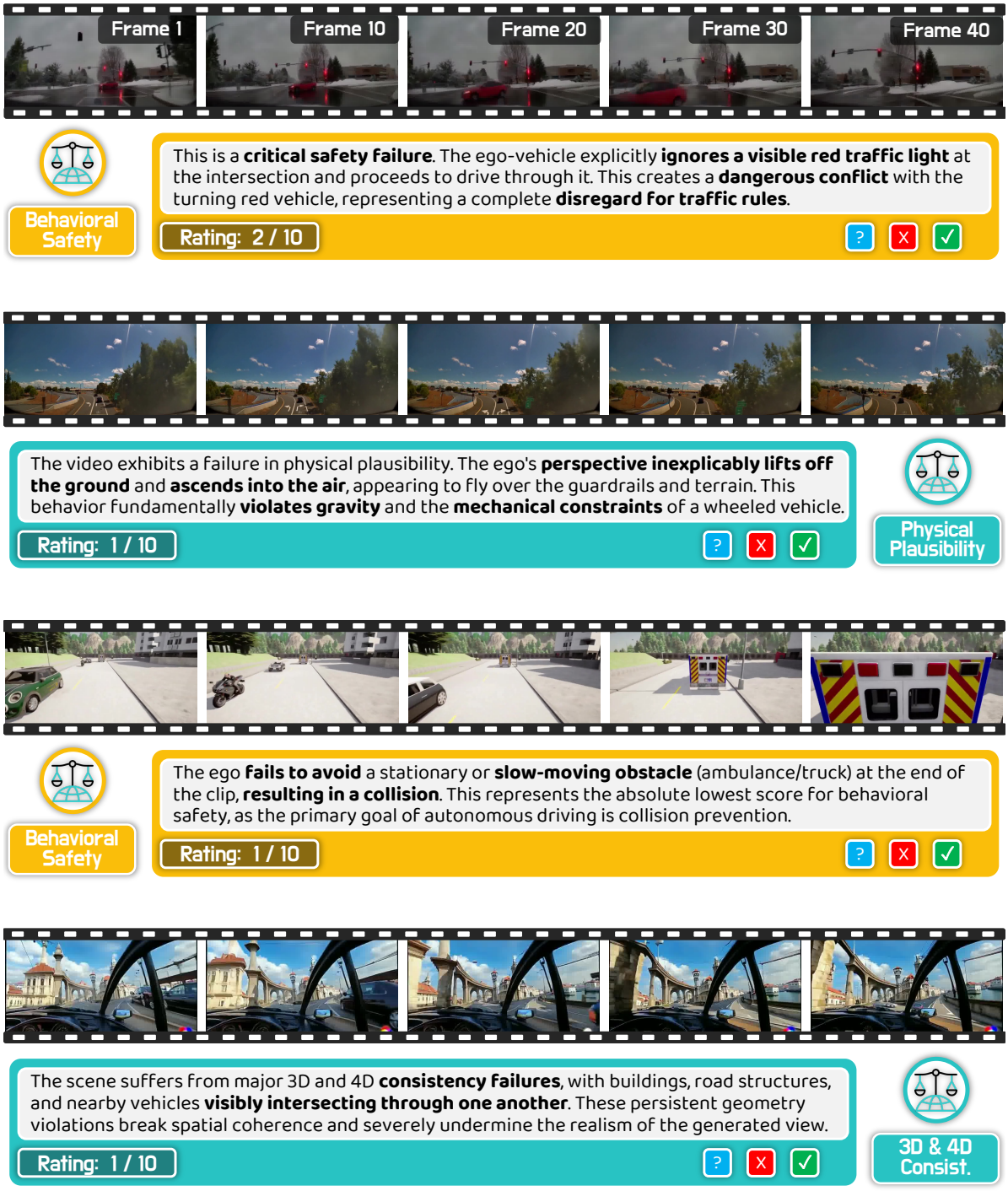


Figure XXIX. Additional qualitative assessments of the **WorldLens-Agent** evaluation on challenging driving conditions.

G. Broader Impact & Limitations

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

G.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.

G.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.

G.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.

H. 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
- DreamForge[6] ..... Apache License 2.0
- DriveDreamer-2[7] ..... Apache License 2.0
- OpenDWM[8] ..... MIT License
- DiST-4D[9] ..... None

[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.
[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.

- $\mathcal{X}$-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
- 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

[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.
[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.

References

[1] 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. 54

[2] 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. 18, 36, 38

[3] 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. 6

[4] 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. 8

[5] 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, 20

[6] 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. 2

[7] 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*, pages 28706–28719, 2024. 28, 30

[8] 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. 2

[9] 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. 38

[10] 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. 55

[11] 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

[12] 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. 26

[13] 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. 14

[14] 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. 22

[15] 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. 36

[16] 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. 34, 36

[17] 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. 18

[18] 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. 2

[19] 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. 10

[20] 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. 20

[21] 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. 55

[22] 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. 55

[23] 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. 16

[24] 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. 14

[25] 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. 40

[26] 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. 18

[27] 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. 2

[28] 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 *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1255–1262. IEEE, 2023. 28

[29] 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. 2

[30] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video generation using VQ-VAE and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 14

[31] 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. 28, 30, 32

[32] 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. 20

[33] 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. 12

[34] 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. 18

[35] 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. 2

[36] 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. 4