MIRAGEDRIVE: SEMANTICALLY ALIGNED APPEARANCE TRANSFER WITH AN EXEMPLAR FOR MULTI-VIEW DRIVING SCENE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Controllable driving-scene generation provides scalable annotated multi-view data for autonomous driving, but its appearance diversity is often constrained by the training distribution. Expanding this diversity requires incorporating new appearance cues while preserving scene geometry and annotation alignment. We introduce **MirageDrive**, an exemplar-guided framework that transfers the appearance of a single reference image to multi-view driving scenes under geometric and semantic conditions. Semantic Region Control supplies dense target-layout guidance, while Exemplar-Guided Appearance Transfer combines class-wise AdaIN pseudo-target supervision with Region-to-Region Attention. The pseudo targets introduce exemplar-derived appearance statistics while retaining the target layout, providing spatially aligned supervision for jointly learning structural control and appearance transfer. Region-to-Region Attention directs reference features to matching semantic regions, coordinating appearance guidance with the target structure. Once trained, MirageDrive accepts new exemplars without additional model adaptation. Experiments on nuScenes demonstrate that MirageDrive achieves semantically aligned appearance transfer, improves controllability, and benefits downstream vision tasks through its generated data.

1 INTRODUCTION

Generative models are becoming an important source of training data for autonomous driving. By conditioning on scene geometry, they generate multi-view images aligned with scene annotations, enabling synthetic data augmentation for downstream tasks (Gao et al., 2024b; Wen et al., 2024). However, the limited geographic coverage of driving datasets can restrict the diversity of environments and appearances available for training (Yang et al., 2024a). As illustrated in Fig. 1(a), a generator trained on Boston driving scenes may struggle to reproduce a Tokyo-like appearance without additional local data and adaptation. Repeated data collection and model adaptation across locations can be costly and time-consuming. An efficient way to expand appearance diversity while retaining control over scene content could provide more varied training data and help downstream models generalize to unfamiliar domains (Yue et al., 2019).

Studies of human perception suggest that people can extrapolate observed visual styles to unseen objects (Boger & Firestone, 2025). Inspired by this ability, we ask **whether a single reference image, as an *exemplar*, can guide the appearance of multi-view driving scenes while preserving control over scene content, without additional training for each new reference**. Realizing this goal requires jointly addressing two challenges. *(i) Preserving structure during appearance transfer:* The generator must incorporate appearance cues from the exemplar without altering the target road layouts, object positions, or geometric relationships across camera views. Otherwise, generated images may no longer match the existing annotations, limiting their use for downstream training and evaluation (Gao et al., 2024b; Wen et al., 2024). *(ii) Aligning appearance within the target structure:* The exemplar and target may differ in spatial layout, requiring appearance cues to be matched by semantic class while respecting the target region boundaries. Colors and textures should be transferred to semantically corresponding regions. Cross-class mixing may reduce the realism of the generated scene (Cherian & Sullivan, 2019).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

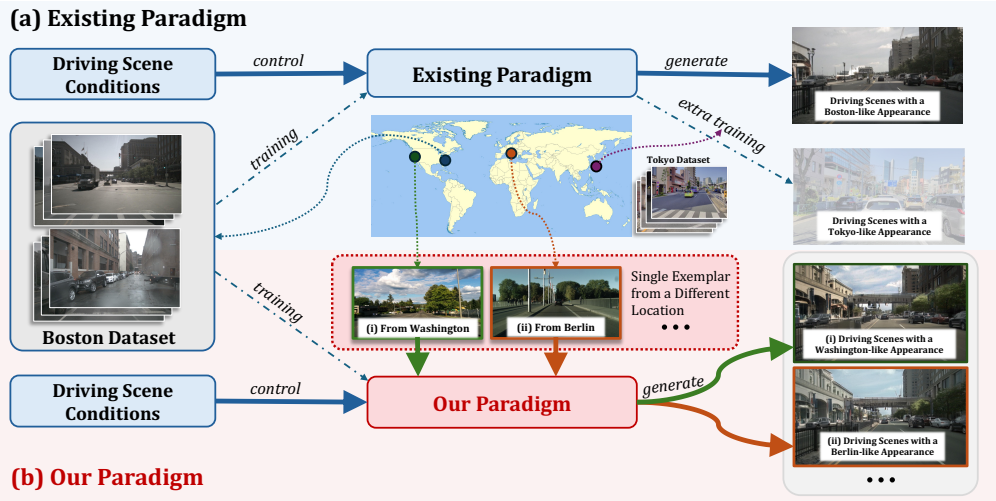

Figure 1: **Generating driving scenes with the appearance of different locations using a single exemplar. (a) Existing paradigm:** a generation model trained on Boston driving data primarily generates scenes with a Boston-like appearance. Adapting it to Tokyo typically requires local driving data and additional training. **(b) Our paradigm:** Under the same driving conditions, our paradigm uses a single exemplar from a different location, such as (i) Washington or (ii) Berlin, to generate scenes with the exemplar's appearance while preserving the target scene geometry.

Existing controllable driving-scene generators (Wen et al., 2024; Mei et al., 2024; Zhao et al., 2024; Gao et al., 2024b) use conditions such as BEV layouts, 3D bounding boxes, perspective guidance, and trajectories to constrain scene geometry. However, these conditions do not explicitly specify the desired colors and textures of semantic regions such as building facades, vegetation, and sky. Text prompts offer additional control over broad attributes, including weather and time of day, but are less suited to precisely specifying the regional appearance. Reference-based style transfer highlights the need to coordinate appearance fidelity with content preservation. However, related methods (Chung et al., 2024; Xu et al., 2025; Chigot et al., 2025; Yao et al., 2026) implement this coordination through inference-time manipulation of pretrained features and latent statistics, without explicitly training the generator on exemplar-derived targets that preserve the conditions. This motivates jointly learning spatial control and appearance transfer through supervision that associates reference appearance with the prescribed scene structure.

In this work, we propose **MirageDrive** (**M**ulti-View **I**mage Generation via **R**egion-**A**ligned **G**uidance from an **E**xemplar), a framework that coordinates appearance transfer from a single exemplar with structural control for multi-view driving-scene generation, as shown in Fig. 1(b). MirageDrive comprises two complementary components. *Semantic Region Control* encodes target semantic maps into dense spatial features and injects them into the control branch to constrain the spatial layout of each generated view. *Exemplar-Guided Appearance Transfer* combines class-wise AdaIN pseudo-target supervision with Region-to-Region (R2R) Attention. The pseudo targets provide reference-derived appearance supervision while retaining the target layout, and R2R Attention restricts reference-feature interaction to matching semantic classes. Together, these components control where appearance changes occur and which reference regions guide them, enabling appearance variation while maintaining alignment with the target geometry and annotations. At inference, the trained model uses a single exemplar to guide scene appearance under the given geometric and semantic conditions, without collecting a new driving dataset or retraining the model for each new exemplar. Our contributions are summarized as follows:

1) We propose MirageDrive, an exemplar-guided framework that transfers appearance from a single exemplar to multi-view driving scenes while maintaining alignment with the target geometry and annotations. Once trained, it supports new exemplars without additional model adaptation.

2) We develop Semantic Region Control to preserve the target spatial layout and Exemplar-Guided Appearance Transfer to transfer appearance between semantically corresponding regions. Together, they align exemplar appearance with the semantic structure of the target scenes.

3) Qualitative comparisons and user studies demonstrate that MirageDrive effectively transfers exemplar appearance to corresponding semantic regions while preserving the target scene layout. Experiments on nuScenes further show improved generation quality and annotation controllability over the matched baseline, together with downstream 3D detection gains from the generated data.

2 Related Work

Controllable driving-scene generation. Controllable driving-scene generators synthesize multi-view images and videos conditioned on scene geometry. BEVGen (Swerdlow et al., 2023) and BEVControl (Yang et al., 2023) use bird's-eye-view layouts. MagicDrive, DrivingDiffusion, and WoVoGen combine geometric scene representations with mechanisms for multi-view consistency (Gao et al., 2024b; Li et al., 2023; Lu et al., 2023). Panacea and DreamForge emphasize temporal coherence in driving-video generation, while MagicDrive-V2 supports high-resolution, long-video generation with spatiotemporal control (Wen et al., 2024; Mei et al., 2024; Gao et al., 2025). UniScene and ConsisDrive further explore occupancy-centric generation and temporal instance consistency, respectively (Li et al., 2025; Yang & Zhang, 2026). These advances improve alignment with driving-scene geometry, but geometric conditions alone do not specify the colors and textures of buildings, vegetation, or sky. Text descriptions can guide appearance, yet may not capture the specific visual details of an exemplar. Our study complements these controls with exemplar guidance aligned to the target semantic regions, enabling appearance variation under the specified scene geometry.

Synthetic-to-real adaptation. Synthetic driving datasets such as SYNTHIA and GTA5 contain images rendered in virtual environments, with corresponding annotations for perception tasks (Ros et al., 2016; Richter et al., 2016). However, differences between rendered and real-world images limit the transfer of models trained on synthetic data. CycleGAN learns mappings between unpaired image domains (Zhu et al., 2017), while CyCADA combines image- and feature-space alignment with semantic consistency for domain adaptation (Hoffman et al., 2018). In synthetic-to-real adaptation, the objective is to reduce the distribution gap between synthetic training data and a real target domain. Our study focuses on exemplar-conditioned appearance variation under specified scene geometry. After training, it accepts new exemplars to guide appearance without further model adaptation, while retaining the target spatial conditions for annotation alignment.

Style and appearance transfer. Early neural style-transfer methods model style through feature correlations or channel statistics (Gatys et al., 2016; Johnson et al., 2016; Huang & Belongie, 2017; Li et al., 2017). Spatial guidance further restricts style transfer to corresponding image regions (Gatys et al., 2017). Diffusion-based approaches support image conditioning through IP-Adapter (Ye et al., 2023), style consistency through shared attention in StyleAligned (Hertz et al., 2023), and reference-based stylization through selective feature injection in InstantStyle or learned style representations in StyleShot (Wang et al., 2024; Gao et al., 2024a). However, appearance guidance alone does not ensure that transferred features remain within the appropriate semantic regions. For driving scenes, appearance transfer must also preserve the spatial structure associated with existing annotations. Our study uses semantic correspondence both to construct exemplar-derived appearance supervision and to restrict exemplar-feature interaction to matching semantic classes. This couples regional appearance transfer with generation conditioned on driving-scene geometry.

3 Method

We propose MirageDrive, a framework for transferring the appearance of a single exemplar to multi-view driving scenes while preserving the target geometry. Semantic Region Control provides spatial guidance through target semantic maps (Sec. 3.2). Exemplar-Guided Appearance Transfer combines Region-to-Region (R2R) Attention for class-matched feature transfer (Sec. 3.3) with class-wise AdaIN pseudo-target supervision (Sec. 3.4). Figure 2 illustrates the overall architecture.

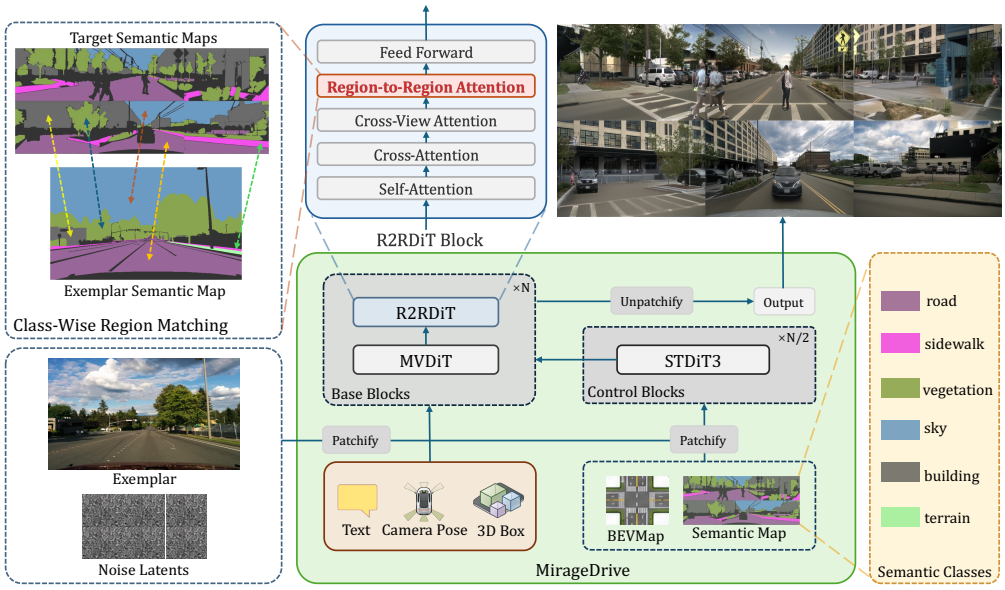

Figure 2: **Overview of MirageDrive.** A single exemplar guides the generation of six target views from noise latents under the given driving conditions. *Semantic Region Control* combines target semantic-map and BEV-map features in the control branch to preserve the target layout. *Exemplar-Guided Appearance Transfer* matches target and exemplar regions by semantic class, allowing R2R Attention in the R2RDiT blocks to transfer features between matching regions.

3.1 Overview

Building on MagicDrive-V2 (Gao et al., 2025), MirageDrive retains its multi-view diffusion-transformer backbone, spatial control branch, and existing generation conditions. The backbone and spatial control blocks are based on STDiT3 from Open-Sora (Zheng et al., 2024).

We refer to the nuScenes scene whose geometry is to be preserved as the target scene and its six camera views as the target views. Using SAM 3 (Carion et al., 2026), we precompute target semantic maps for the camera images in the nuScenes training and validation sets. A single exemplar provides the desired appearance for all six target views.

Semantic Region Control encodes the target semantic maps and combines their features with the BEV-map features in the shared control branch, providing spatial guidance for generation. For Exemplar-Guided Appearance Transfer, the exemplar is encoded by the CogVideoX 3D VAE (Yang et al., 2024b) and converted into tokens. R2R Attention allows target views to access these tokens within matching semantic classes. During training, class-wise AdaIN constructs pseudo targets that combine exemplar appearance with the target structure.

3.2 Semantic Region Control

BEV maps and 3D boxes describe scene geometry but provide limited image-plane guidance for appearance transfer. Semantic Region Control uses target semantic maps to specify the locations and semantic classes of regions in the generated target views.

The target and exemplar semantic maps share six static classes: road, sidewalk, building, vegetation, terrain, and sky. We encode each class name L_c using T5 (Raffel et al., 2020), project the token embeddings to the model dimension, and average them to obtain a class embedding e_c:

$$e_c = \mathrm{AvgPool}(E_{\mathrm{text}}(L_c)) \,. \tag{1}$$

Here, E_{text} includes T5 encoding and the model's text projection. For each target view i, we fill each region of class c with its corresponding embedding e_c, forming a semantic feature map F_i. A lightweight semantic encoder converts this map into features that are patchified and combined with

the BEV-map tokens:

$$T_i^{\text{sem}} = \text{Patchify}(E_{\text{sem}}(F_i)), \qquad T_i^{\text{cond}} = T_i^{\text{map}} + T_i^{\text{sem}}. \tag{2}$$

Here, E_{sem} is a lightweight four-layer CNN (LeCun et al., 1998), and T_i^{map} denotes the encoded and projected BEV-map tokens. The combined tokens enter the shared control branch (Zhang et al., 2023), providing geometric and semantic guidance for generation.

3.3 Region-to-Region Attention

R2R Attention transfers appearance features between target and exemplar regions with the same semantic label. It operates between cross-view attention and the feed-forward layer.

The exemplar latent is converted into tokens using the same patch-embedding layer as the target latents. For each target view, we resize the semantic masks to the token grid and group target and exemplar tokens by class. Let Q_c, K_c, and V_c denote the queries, keys, and values of the target region of class c. When the exemplar contains the same class, its tokens provide additional keys K_c^{exemplar} and values V_c^{exemplar}. Otherwise, attention uses only the target region's keys and values, reducing to self-attention within that region:

$$(\overline{K}_c, \overline{V}_c) = \begin{cases} \big([K_c^{\text{exemplar}}, K_c],\, [V_c^{\text{exemplar}}, V_c]\big), & \text{if class } c \text{ is present in the exemplar,} \\ (K_c, V_c), & \text{otherwise.} \end{cases} \tag{3}$$

Here, square brackets denote concatenation along the token dimension. Attention is computed as

$$A_c = \text{softmax}\left(\frac{Q_c \overline{K}_c^{\top}}{\sqrt{d_h}}\right)\overline{V}_c, \tag{4}$$

where d_h is the dimension of each attention head. Access to both target and exemplar features helps each region preserve its own structure while incorporating exemplar appearance, without attending to unrelated semantic classes.

3.4 Class-wise AdaIN Pseudo Targets and Training Objective

An exemplar provides the desired appearance but does not share the target scene layout. We therefore construct pseudo targets that combine exemplar appearance with the structure of the target-view ground-truth images.

For each target view, we identify semantic classes present in both the ground-truth image and the exemplar using their semantic masks. For each shared class c, let x_c and E_c denote the RGB pixels in the target and exemplar regions, respectively. We apply AdaIN (Huang & Belongie, 2017) only to these matching semantic regions:

$$\widehat{x}_c = \sigma(E_c)\frac{x_c - \mu(x_c)}{\sigma(x_c)} + \mu(E_c), \tag{5}$$

where μ and σ are computed separately for each RGB channel over the pixels in each region. This transformation transfers the exemplar region's color statistics to the target region of the same class.

We then blend the transformed region $\widehat{x}_c$ with the original target region x_c:

$$\widetilde{x}_c = (1-\lambda)x_c + \lambda\widehat{x}_c, \qquad \lambda \in [0,1]. \tag{6}$$

Here, λ controls the strength of appearance transfer. Blending helps retain target image details and reduces structural degradation caused by strong appearance changes. Each blended region replaces only the corresponding semantic region in the target-view ground-truth image. For example, transformed building pixels replace the original pixels only within the target building region. Target regions without a matching semantic class in the exemplar remain unchanged. Applying this process to all target views produces the pseudo target $\widetilde{x}$.

We encode $\widetilde{x}$ with the VAE to obtain its latent z_1. For flow-matching training (Lipman et al., 2023), we interpolate between Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ and the pseudo-target latent:

$$z_t = (1-t)\epsilon + tz_1, \qquad t \in [0,1]. \tag{7}$$

The model learns to predict the velocity from noise to the pseudo-target latent:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{\epsilon,t}\left[\|v_\theta(z_t, t, y) - (z_1 - \epsilon)\|_2^2\right], \tag{8}$$

where v_θ is the predicted velocity and y includes the driving conditions, target semantic maps, and exemplar inputs.

AdaIN pseudo targets define the appearance changes the model should learn. R2R Attention provides region-level control by restricting access to exemplar features to semantically matching regions. At inference, the exemplar directly guides generation from Gaussian noise without constructing pseudo targets.

4 Experiments

4.1 Experimental setup

Datasets and baselines. We use the nuScenes dataset (Caesar et al., 2020), comprising 700 training scenes and 150 validation scenes, with six synchronized camera views and corresponding driving annotations per sample. Exemplars are sampled from Cityscapes (Cordts et al., 2016) and Mapillary Vistas (Neuhold et al., 2017). Cityscapes provides 5,000 finely annotated images from 50 cities, while Mapillary Vistas contains 25,000 annotated street-level images spanning six continents. These datasets provide diverse scene appearances, and their semantic labels are mapped to the six classes used by MirageDrive.

Model training uses exemplars only from the Cityscapes and Mapillary Vistas training sets. Data generation uses their validation sets, with 350 exemplars for training data and 75 for validation data, at one exemplar per two target scenes. The 75 exemplars include 15 Cityscapes and 60 Mapillary Vistas images, randomly sampled in proportion to validation-set sizes. Each target scene uses a fixed exemplar across all six views. Generated validation data assess generation quality and controllability, while generated training data assess downstream perception performance with StreamPETR.

We compare MirageDrive with representative driving-scene generation methods, including BEVGen (Swerdlow et al., 2023), BEVControl (Yang et al., 2023), MagicDrive (Gao et al., 2024b), Panacea (Wen et al., 2024), and MagicDrive-V2 (Gao et al., 2025).

Evaluation metrics. We use Fréchet Inception Distance (FID) to measure image fidelity (Heusel et al., 2017). For controllability, pretrained CVT (Zhou & Krähenbühl, 2022) evaluates road and vehicle segmentation using mean Intersection over Union (mIoU), while camera-only BEVFusion (Liu et al., 2023) evaluates 3D detection using mean Average Precision (mAP) and nuScenes Detection Score (NDS). These perception metrics measure how well generated images retain the conditioned scene layout and object annotations. Specifically, mAP measures detection accuracy, whereas NDS combines detection accuracy with localization, size, orientation, velocity, and attribute errors (Caesar et al., 2020). For downstream StreamPETR experiments, we report mAP, NDS, and mean Average Orientation Error (mAOE), which measures object orientation error. We compare the appearance of generated images and their exemplars using Gram matrices of VGG19 features and report the average style distance (Style dist.) across all samples and camera views.

Implementation details. All experiments use images at 424×800 resolution unless otherwise specified. We initialize MirageDrive from MagicDrive-V2 (Gao et al., 2025) and train for 18,750 iterations on four NVIDIA A800 GPUs using the flow-matching objective (Lipman et al., 2023). We use learning rates of 10^{-5} for the R2R Attention and semantic encoding modules, 2×10^{-6} for the spatial control blocks, and 10^{-6} for the last 13 spatial backbone blocks and output layer. We set $\lambda = 0.5$ for AdaIN pseudo targets. We use 10 sampling steps for training-data generation and 30 for validation-data generation, with a classifier-free guidance scale of 2.0. Downstream experiments use StreamPETR with a ResNet-50 backbone and 256×704 image inputs (Wang et al., 2023).

4.2 Generation quality and controllability

We generate target views conditioned on the nuScenes validation annotations and evaluate their fidelity and controllability. Table 1 compares published baseline results with our matched MagicDrive-V2 and MirageDrive evaluations.



Figure 3: Exemplar-guided driving-scene generation on nuScenes. From left to right: ground truth, MagicDrive-V2, MirageDrive, and the exemplar. Across diverse target scenes and exemplars, MirageDrive transfers appearance cues such as sky color, road texture, building appearance, and illumination while preserving the target scene layout.

Table 1: Comparison with selected driving-scene generation methods on nuScenes validation.

Method	FID ↓	Road mIoU ↑	Vehicle mIoU ↑	mAP ↑	NDS ↑
BEVGen (Swerdlow et al., 2023)	25.54	50.20	5.89	–	–
BEVControl (Yang et al., 2023)	24.85	60.80	26.80	–	–
MagicDrive (Gao et al., 2024b)	16.20	61.05	27.01	12.30	23.32
Panacea (Wen et al., 2024)	16.96	55.78	22.74	11.58	22.31
MagicDrive-V2 (Gao et al., 2025)	17.23	58.75	31.80	13.72	23.92
MirageDrive	**15.71**	**62.42**	**33.02**	**17.95**	**29.13**

Compared with MagicDrive-V2, MirageDrive reduces FID from 17.23 to 15.71 and improves detection mAP from 13.72 to 17.95. NDS increases from 23.92 to 29.13. Road mIoU improves from 58.75 to 62.42, and vehicle mIoU increases from 31.80 to 33.02. These results show improved image fidelity and stronger alignment with the conditioned scene layout and object annotations.

4.3 Qualitative Analysis

As shown in Figure 3, MirageDrive transfers exemplar appearance while preserving target structure. Compared with MagicDrive-V2, it better retains building layouts, road geometry, and roadside structures in (a)–(e), particularly the number of roadside poles on the right in (a) and the streetlight structure in (b). Appearance changes reflect the exemplars' darker tones in (a) and (e) and warmer road and building colors in (c). In (d), the skybridge is preserved alongside the exemplar's brighter blue sky and clouds. In the low-light scene (f), sky brightness and color shift toward the exemplar while the target road direction is maintained. These examples show that MirageDrive transfers exemplar appearance while preserving the target scene structure. Semantic Region Control guides the positions and boundaries of target semantic regions. AdaIN pseudo-target supervision defines

the appearance changes the model should learn, while R2R Attention transfers exemplar features between matching semantic regions.

4.4 HUMAN EVALUATION

We recruited 20 master's and doctoral students conducting research in autonomous driving for a questionnaire study using 50 examples, each associated with a different exemplar randomly selected from the validation exemplar pool. For each exemplar, we selected one corresponding generated sample and presented its six target views alongside the ground-truth views and the exemplar.

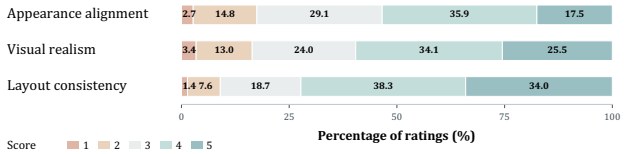

Figure 4: Human evaluation on 50 generated multi-view samples. Bars show the percentage of valid ratings at each score, from 1 (lowest) to 5 (highest).

Respondents assessed all six views together and rated each example from 1 to 5 on three criteria: 1) **Appearance alignment:** How closely do the colors and textures of the generated scene match the corresponding semantic regions in the exemplar? 2) **Visual realism:** How natural and realistic does the generated scene appear? 3) **Layout consistency:** How well does the generated scene preserve the road layout, building structure, and object positions shown in the ground-truth views?

We collected 20 completed questionnaires. As shown in Figure 4, ratings of 4 or 5 accounted for 53.4%, 59.6%, and 72.3% of responses for appearance alignment, visual realism, and layout consistency, respectively. The corresponding mean ratings were 3.51, 3.65, and 3.96, with sample variances of 1.06, 1.20, and 0.96, respectively. Layout consistency received the highest proportion of high ratings. These ratings provide perceptual support for appearance transfer, visual realism, and target-layout preservation.

4.5 ABLATION STUDY

We examine Semantic Region Control, Region-to-Region (R2R) Attention, and class-wise AdaIN pseudo-target supervision through qualitative and quantitative comparisons. Qualitative results are shown in Fig. 5, and quantitative results are reported in Tab. 2. We progressively remove Semantic Region Control and R2R Attention from the full model and include MagicDrive-V2 as the baseline.

Table 2: Cumulative ablation study on nuScenes validation. Components are removed cumulatively in the first three rows.

Ablation	FID ↓	Road mIoU ↑	Vehicle mIoU ↑	mAP ↑	Style dist. ↓
Full MirageDrive	15.71	62.42	33.02	17.95	0.9796
– Semantic Region Control	16.77	60.92	33.05	17.60	0.9821
– R2R Attention	18.06	58.63	32.55	17.04	1.0031
MagicDrive-V2 (Gao et al., 2025) (baseline)	17.23	58.75	31.80	13.72	1.0957



Figure 5: Qualitative comparison of the cumulative ablation settings. From left to right: ground truth, MagicDrive-V2, MirageDrive without Semantic Region Control (SRC) and R2R Attention, MirageDrive without SRC, full MirageDrive, and the exemplar. Red boxes highlight differences in the skybridge connecting the buildings.

Semantic Region Control. To assess the contribution of dense semantic guidance, we first remove Semantic Region Control from the full model. As shown in Fig. 5, the skybridge connecting the buildings disappears without this component, whereas the full model preserves this structure. In Tab. 2, removing Semantic Region Control increases FID from 15.71 to 16.77 and reduces road mIoU from 62.42 to 60.92 and mAP from 17.95 to 17.60. These results support the role of dense semantic guidance in preserving target structure during appearance transfer.

Region-to-Region Attention. We further remove R2R Attention to examine its contribution to class-matched appearance transfer. As shown in Fig. 5, the buildings, sky, and road exhibit more muted colors, with weaker resemblance to the exemplar. Quantitatively, style distance increases from 0.9821 to 1.0031, and FID increases from 16.77 to 18.06. Road mIoU, vehicle mIoU, and mAP decrease by 2.29, 0.50, and 0.56 points, respectively. Even without SRC, R2R Attention improves exemplar similarity and scene controllability through class-matched feature transfer.

Class-wise AdaIN Pseudo-Target Supervision. With both Semantic Region Control and R2R Attention removed, the model still uses class-wise AdaIN pseudo-target supervision. We compare this setting with the MagicDrive-V2 baseline. In Fig. 5, the model trained with pseudo targets exhibits changes in scene coloration, but does not reproduce the exemplar appearance as clearly as the full model. In Tab. 2, it achieves 17.04 mAP and 32.55 vehicle mIoU, compared with 13.72 and 31.80 for MagicDrive-V2. However, FID increases from 17.23 to 18.06, while road mIoU remains similar. This setting changes scene coloration and improves detection controllability, but does not improve all generation metrics.

4.6 Downstream Perception Enhancement

We evaluate the utility of generated data for downstream 3D detection using StreamPETR. Table 3 reports results for training on generated data and for pretraining on generated data followed by fine-tuning on nuScenes training images.

Table 3: Comparison of downstream 3D detection performance on nuScenes validation.

Training data	Data source	mAP ↑	NDS ↑	mAOE ↓
Real	Ori nuScenes	37.45	47.62	64.56
Generated	Panacea (Wen et al., 2024)	22.50	36.10	72.70
	DreamForge (Mei et al., 2024)	26.00	41.10	**62.20**
	MagicDrive-V2 (Gao et al., 2025)	26.66	40.07	69.54
	MirageDrive	**30.11**	**43.25**	64.06
Real + Generated	Panacea (Wen et al., 2024)	37.10	49.20	54.20
	DreamForge (Mei et al., 2024)	36.60	49.10	52.90
	Dream4Drive (Zeng et al., 2026)	38.70	50.60	–
	MagicDrive-V2 (Gao et al., 2025)	39.51	50.68	**52.26**
	MirageDrive	**41.27**	**51.46**	53.23

Training on generated data. StreamPETR trained on MirageDrive data achieves 30.11 mAP and 43.25 NDS, versus 26.66 and 40.07 with MagicDrive-V2 data, while reducing mAOE from 69.54 to 64.06. These gains show that MirageDrive data provide useful supervision for 3D detection on real scenes.

Training with real and generated data. After fine-tuning on nuScenes, StreamPETR pretrained on MirageDrive data achieves 41.27 mAP and 51.46 NDS, compared with 39.51 mAP and 50.68 NDS when pretrained on MagicDrive-V2 data. These results also exceed the real-data-only baseline of 37.45 mAP and 47.62 NDS. These sustained gains support the value of MirageDrive data for detector pretraining.

5 Conclusion

We presented MirageDrive, an exemplar-guided framework for varying the appearance of multi-view driving scenes while preserving target geometry and annotation alignment. In MirageDrive, Semantic Region Control anchors the target layout, while Exemplar-Guided Appearance Transfer combines AdaIN pseudo-target supervision and Region-to-Region Attention for class-matched appearance transfer. Results on nuScenes demonstrate that appearance diversification can preserve annotation alignment and provide effective training data for downstream vision tasks. Future work will extend this framework to long driving videos, focusing on maintaining consistent exemplar appearance and scene structure over time.

AI USE STATEMENT

Generative AI tools assisted with translation, language polishing, terminology refinement, and discussions of mathematical formulations during manuscript preparation. The core research ideas, methodological design, and original contributions were developed by the authors. The authors take responsibility for the final manuscript and results.

ETHICS STATEMENT

This work studies driving-scene generation using publicly available datasets, including nuScenes, Cityscapes, and Mapillary Vistas. The human evaluation involved 20 master's and doctoral students conducting autonomous-driving research, with informed consent obtained before participation. All questionnaires were anonymous and identified only by numerical codes. Only aggregate ratings are reported. The generated data are intended for research and require further safety validation before use in real-world driving systems.

REPRODUCIBILITY STATEMENT

Section 3 describes the model architecture and training objective. Section 4.1 specifies the datasets, exemplar selection, training settings, and evaluation metrics, while Section 4.4 describes the human-evaluation procedure. Appendix A provides additional training and evaluation details. All source datasets used in this work are publicly available.

REFERENCES

Tal Boger and Chaz Firestone. The psychophysics of style. *Nature Human Behaviour*, 9:2497–2509, 2025. doi: 10.1038/s41562-025-02249-8.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026.

Anoop Cherian and Alan Sullivan. Sem-GAN: Semantically-Consistent Image-to-Image Translation. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 1797–1806, 2019. doi: 10.1109/wacv.2019.00196. URL https://doi.org/10.1109/wacv.2019.00196.

Estelle Chigot, Dennis G. Wilson, Meriem Ghrib, and Thomas Oberlin. Style transfer with diffusion models for synthetic-to-real domain adaptation. *Computer Vision and Image Understanding*, 2025. doi: 10.1016/j.cviu.2025.104445. URL https://arxiv.org/abs/2505.16360.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer. *arXiv preprint arXiv:2312.09008*, 2024. URL https://arxiv.org/abs/2312.09008.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *arXiv preprint arXiv:2407.01414*, 2024a.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *International Conference on Learning Representations*, 2024b.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. *arXiv preprint arXiv:2312.02133*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1989–1998. PMLR, 2018. URL https://proceedings.mlr.press/v80/hoffman18a.html.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://bottou.org/papers/lecun-98h.

Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, Shuchang Zhou, Li Zhang, Xiaojuan Qi, Hao Zhao, Mu Yang, Wenjun Zeng, and Xin Jin. Uniscene: Unified occupancy-centric driving scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint arXiv:2310.07771*, 2023.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, 2017.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela 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*, 2023.

Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. *arXiv preprint arXiv:2312.02934*, 2023.

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

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://www.jmlr.org/papers/v21/20-074.html.

Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for Data: Ground Truth from Computer Games. In *European Conference on Computer Vision*, pp. 102–118, 2016. doi: 10.1007/978-3-319-46475-6_7. URL https://doi.org/10.1007/978-3-319-46475-6_7.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3234–3243, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Ros_The_SYNTHIA_Dataset_CVPR_2016_paper.html.

Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *arXiv preprint arXiv:2301.04634*, 2023.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024.

Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. *arXiv preprint arXiv:2303.11926*, 2023.

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and Controllable Video Generation for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6902–6912, June 2024.

Ruojun Xu, Weijie Xi, XiaoDi Wang, Yongbo Mao, and Zach Cheng. StyleSSP: Sampling Start-Point Enhancement for Training-free Diffusion-based Method for Style Transfer. *arXiv preprint arXiv:2501.11319*, 2025. URL https://arxiv.org/abs/2501.11319.

Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized Predictive Model for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14662–14672, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_Generalized_Predictive_Model_for_Autonomous_Driving_CVPR_2024_paper.html.

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

Zhuoran Yang and Yanyong Zhang. Consisdrive: Identity-preserving driving world models for video generation by instance mask. In *International Conference on Learning Representations*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024b.

Dingyi Yao, Xinqi Zhang, Lihui Peng, Jianming Hu, Danya Yao, and Yi Zhang. ASTAD: Asymmetric style transfer for synthetic-to-real adaptation in autonomous driving. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2606.29286.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization without Accessing Target Domain Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2100–2110, 2019. doi: 10.1109/iccv.2019.00219. URL https://doi.org/10.1109/iccv.2019.00219.

Kai Zeng, Zhanqian Wu, Kaixin Xiong, Xiaobao Wei, Xiangyu Guo, Zhenxin Zhu, Kalok Ho, Lijun Zhou, Bohan Zeng, Ming Lu, Haiyang Sun, Bing Wang, Guang Chen, Hangjun Ye, and Wentao Zhang. Rethinking Driving World Model as Synthetic Data Generator for Perception Tasks. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 135507–135534, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/db6caae0f83e45e454e2215f07e7c5af-Paper-Conference.pdf.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing Efficient Video Production for All. *arXiv preprint arXiv:2412.20404*, 2024. doi: 10.48550/arXiv.2412.20404. URL https://arxiv.org/abs/2412.20404.

Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13760–13769, 2022.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2242–2251, 2017. doi: 10.1109/iccv.2017.244. URL https://doi.org/10.1109/iccv.2017.244.

APPENDIX

A ADDITIONAL TRAINING AND EVALUATION DETAILS

Training exemplar sampling. We sample training exemplars exclusively from the Cityscapes and Mapillary Vistas training sets. A cyclic sampling strategy exposes target scenes to different exemplar appearances during training. Each target sample is assigned a starting position in the shuffled exemplar pool based on its sample index. At each epoch, this position advances by a fixed offset, wrapping around at the end of the pool. The same target sample therefore receives different exemplars across epochs rather than keeping a fixed pairing. Each exemplar guides all six target views.

Inference with unseen exemplars. Exemplars used for data generation come from the validation sets of Cityscapes and Mapillary Vistas and are unseen during model training. At inference, a single exemplar guides all six target views without constructing AdaIN pseudo targets or fine-tuning the model. Semantic Region Control specifies the positions and boundaries of target semantic regions, while R2R Attention transfers features from matching exemplar regions. The qualitative results illustrate appearance transfer from unseen exemplars while preserving target scene structure.

Style distance. We use VGG19 Gram-matrix style distance (Style dist.), based on the feature-correlation representation of Gatys et al. (2016). We normalize input images using ImageNet statistics and extract features after the first ReLU activation in each of the first four VGG19 blocks. At each layer, we divide the Gram matrix by the number of spatial positions and compute the mean squared difference between corresponding entries for the generated image and its exemplar. We average these distances across the four layers, then across the six views and all evaluated samples. This whole-image metric measures the similarity of appearance statistics but does not directly evaluate semantic-region correspondence.

B Additional Multi-View Qualitative Results

We present full six-view comparisons for three examples from Figure 3, corresponding to panels (d), (a), and (b), respectively. Each example compares the six nuScenes ground-truth views with MagicDrive-V2 and MirageDrive outputs under the corresponding driving conditions. The exemplar occupies a separate row between the MirageDrive outputs and the target semantic maps extracted using SAM 3. All four six-view groups use the same camera order. These selected examples complement the aggregate evaluation in Section 4.



Figure 6: Full six-view comparison for Figure 3(d): a skybridge scene with a bright-sky exemplar.

In Figure 6, MirageDrive preserves the skybridge connecting the buildings, which is absent from the MagicDrive-V2 output in the front view. The sky becomes bluer and building tones become warmer, reflecting the exemplar appearance while retaining the target road corridor.



Figure 7: Full six-view comparison for Figure 3(a): a rainy target scene with a Cityscapes exemplar.

In Figure 7, MirageDrive transfers the exemplar's darker tones to the road and building regions while preserving the target intersection. Compared with MagicDrive-V2, the front view better retains the building outlines and roadside poles on the right, alongside these appearance changes.

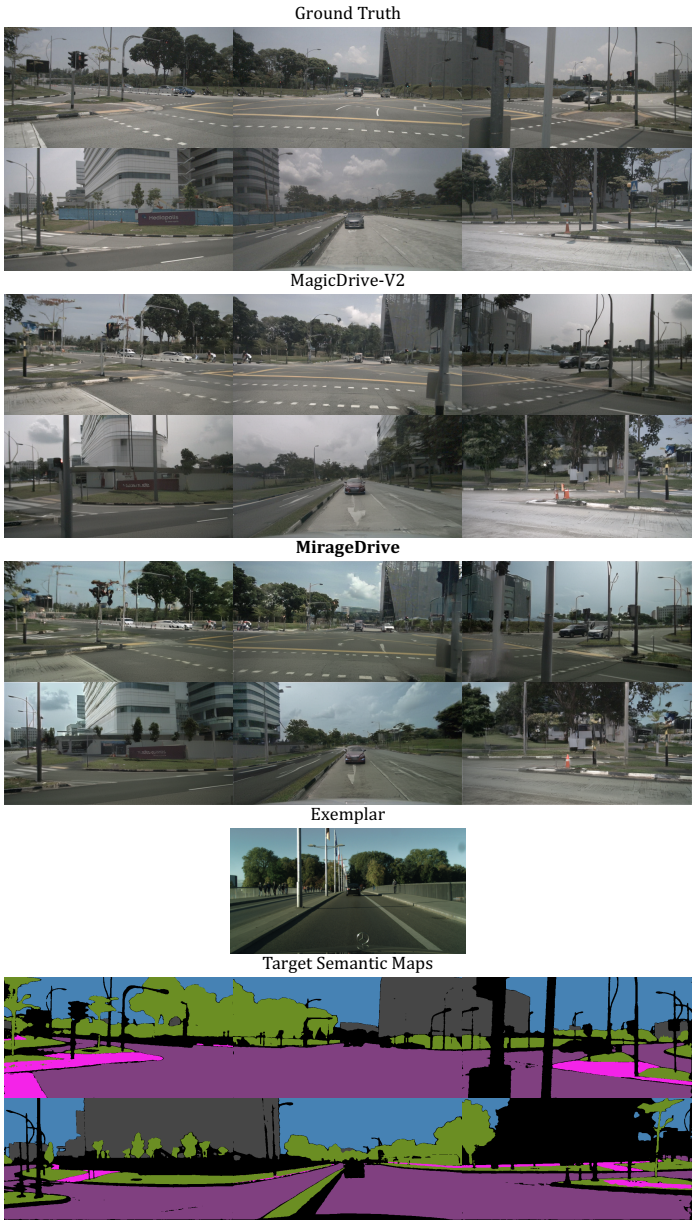

Figure 8: Full six-view comparison for Figure 3(b): an intersection with a clear-sky Cityscapes exemplar.

In Figure 8, MirageDrive retains the curved streetlight structure more completely than MagicDrive-V2 while preserving the intersection layout and surrounding building outlines. The sky and vegetation reflect the exemplar's blue and green tones without changing the target road geometry.

C HUMAN EVALUATION QUESTIONNAIRE

Participants rated each sample from 1 to 5 using the following instructions.

1. Appearance alignment. How closely do the colors and textures of regions such as sky, road, buildings, and vegetation in the generated images match the corresponding regions in the exemplar? Compare regions of the same semantic class. The objects and composition need not be the same. A score of 1 means not at all aligned, and 5 means very well aligned.

2. Visual realism. How much do the generated images resemble photographs of real driving scenes? Consider whether object shapes, textures, and lighting look natural, and whether there are obvious distortions, ghosting, or generation artifacts. A score of 1 means clearly unrealistic, and 5 means very realistic.

3. Layout consistency. How well do the generated images preserve the ground-truth road layout, building outlines, and vehicle positions? Compare corresponding locations in the same camera view. Do not treat changes in color or material appearance as layout changes. A score of 1 means clearly changed, and 5 means highly consistent.

Please rate your actual visual impression. There are no predetermined correct answers. Figure 9 shows only the first six questionnaire items on the left and an enlarged example on the right.

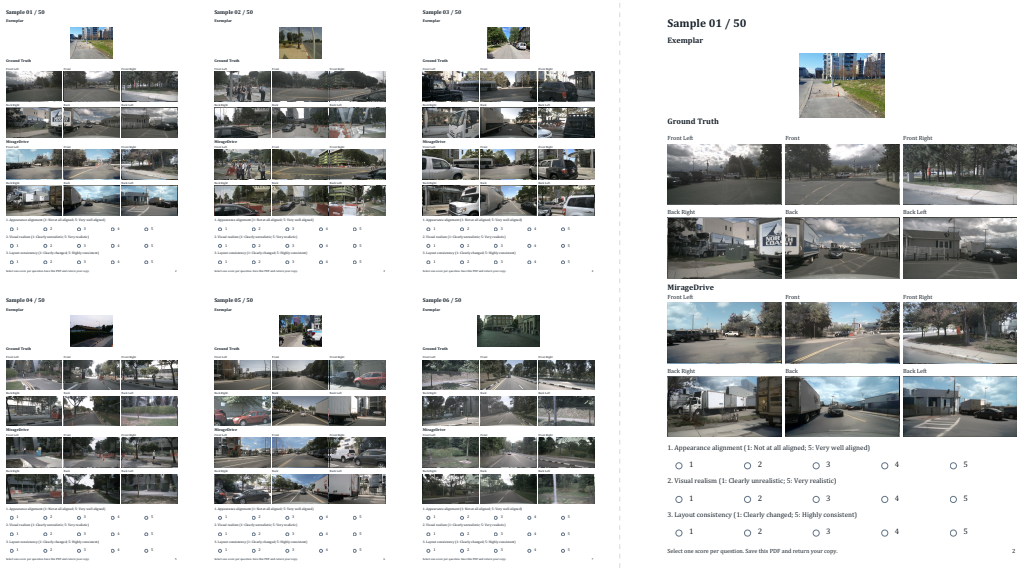

Figure 9: Human evaluation questionnaire. Left: the first six of 50 examples. Right: an enlarged view of the first example, showing the exemplar, six ground-truth views, six MirageDrive-generated views, and rating options.