V-REG: VISUAL REPRESENTATION GUIDANCE FOR TRAINING-EFFICIENT MULTI-VIEW DRIVING VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view driving video generation requires modeling complex temporal, cross-view, and multimodal scene structures, leading to expensive task-specific optimization. We propose V-REG, a pretrained visual representation-guided framework that accelerates the adaptation of a multi-view diffusion transformer while largely preserving its original generation pathway. V-REG introduces frame-view aligned representation entanglement and intermediate feature alignment, associating each pretrained representation with the corresponding frame and camera view. Frozen DINOv3 features serve as both conditioning signals and alignment targets during training; at inference, only one-shot condition encoding is retained, and a KV cache mechanism is employed to accelerate multi-frame autoregressive generation by reusing cached attention states from previous frames. Experiments on nuScenes show that V-REG achieves 8.98 FID and 98.75 FVD after 30K task-specific optimization steps. At the matched FID of 11.3, it requires approximately 4.5 times fewer optimization iterations than the corresponding baseline.

1 INTRODUCTION

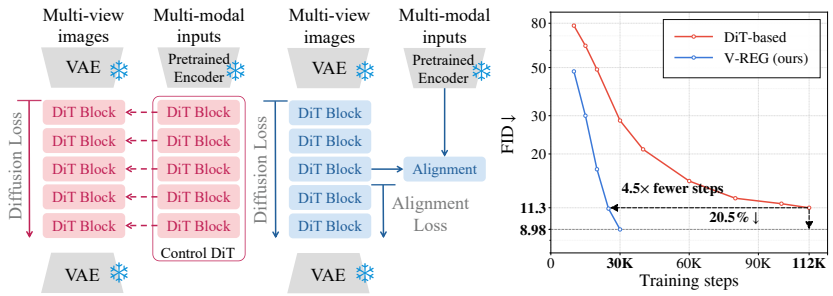
Multi-view driving video generation (Li et al., 2024b) aims to synthesize temporally continuous and geometrically consistent driving scenes from surrounding cameras under controllable conditions, such as camera poses (Gao et al., 2024a), ego trajectories (Gui et al., 2026), semantic layouts (Li et al., 2025), depth cues (Liang et al., 2025), and reference observations. Unlike general image and video generation (Peebles & Xie, 2023; Ma et al., 2024), this task must jointly model realistic scene appearance, temporal dynamics, and cross-view correspondence within a shared 3D environment, making it important for scalable autonomous-driving simulation.

Despite recent progress, training multi-view driving video generators remains computationally expensive. Large generative backbones must learn complex scene semantics, temporal evolution, and multi-camera dependencies primarily through the denoising objective, often resulting in slow convergence and prolonged optimization. Moreover, increasing training cost does not always yield proportional improvements in visual fidelity or temporal and cross-view consistency, revealing an unfavorable efficiency–quality trade-off.

Existing methods introduce task-specific perception priors to improve structured scene generation. For example, UniFuture (Liang et al., 2025) jointly models RGB and depth to constrain scene geometry, while OccScene (Li et al., 2025) incorporates semantic occupancy features to provide geometric and semantic guidance. Although effective, these approaches typically require additional prediction branches, cross-modal interaction modules, or dense supervision. Such designs increase optimization complexity but do not directly reduce the representation-learning burden of the original generative backbone.

In contrast, we leverage pretrained visual representations (Leng et al., 2025) as reusable semantic and structural priors. Since pretrained visual encoders already capture rich scene-level regularities, transferring their representations to the diffusion backbone can reduce the need to relearn such knowledge solely from denoising supervision. However, conventional image-level representation guidance assumes a single global correspondence between one generated sample and one visual representation. This assumption is insufficient for synchronized multi-view videos, where each rep-

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.



(a) Multi-modal DiT Training (b) V-REG (Ours) (c) Training Steps vs. FID Improved Convergence

Figure 1: **Can multimodal representation alignment accelerate multi-view diffusion training?** — Multi-modal DiT introduces an auxiliary Control DiT to inject features from a frozen pretrained encoder, but this additional branch increases training and computational overhead (a). In contrast, V-REG directly aligns intermediate DiT representations with pretrained multimodal features through a lightweight objective, eliminating the extra Control DiT while preserving the original diffusion pipeline (b). Notably, this allows for significantly accelerated training; speeding up training by over 4.5× over MMDiT training recipes, respectively (c).

resentation must remain associated with a specific temporal index and camera viewpoint. Global aggregation discards this structured correspondence and may therefore weaken temporal evolution and cross-view consistency.

To address this issue, we propose V-REG, a correspondence-preserving representation guidance framework for training-efficient multi-view driving video generation. Instead of pooling all video representations into a single global token, V-REG maintains an independent representation for each frame-view pair during both representation entanglement and intermediate feature alignment. This design transfers pretrained semantic and structural knowledge while preserving temporal and camera-view correspondence.

As shown in Figure 1(c), V-REG substantially accelerates convergence while maintaining competitive generation quality. Experiments on nuScenes demonstrate strong performance in RGB video and panoramic depth generation, while the generated multimodal sequences also support competitive downstream 3D semantic occupancy reconstruction.

Our main contributions are summarized as follows:

- We propose V-REG, a pretrained visual representation-guided framework that reduces the task-specific optimization burden of multi-view driving video diffusion without introducing an auxiliary Control DiT.
- We develop frame-view aligned representation entanglement and alignment, preserving the correspondence among pretrained features, video frames, and camera views throughout representation transfer.
- Under controlled training settings, V-REG consistently improves iteration-level convergence while achieving competitive RGB video and panoramic depth generation, together with reliable downstream semantic occupancy reconstruction.

2 Related Works

2.1 World Models for Autonomous Driving

Representative driving world models explore structured scene generation from complementary perspectives. UniFuture (Liang et al., 2025) jointly models future RGB and depth in a shared latent space, while WorldDrive (Gui et al., 2026) unifies vision and motion representations for trajectory-aware generation and downstream planning. FAR-Drive (Li et al., 2026) further adopts frame-level autoregressive generation for long-horizon closed-loop simulation. Overall, these methods highlight the importance of geometry, temporal consistency, and planning awareness, but often depend on additional reconstruction, autoregressive, or task-specific modules.

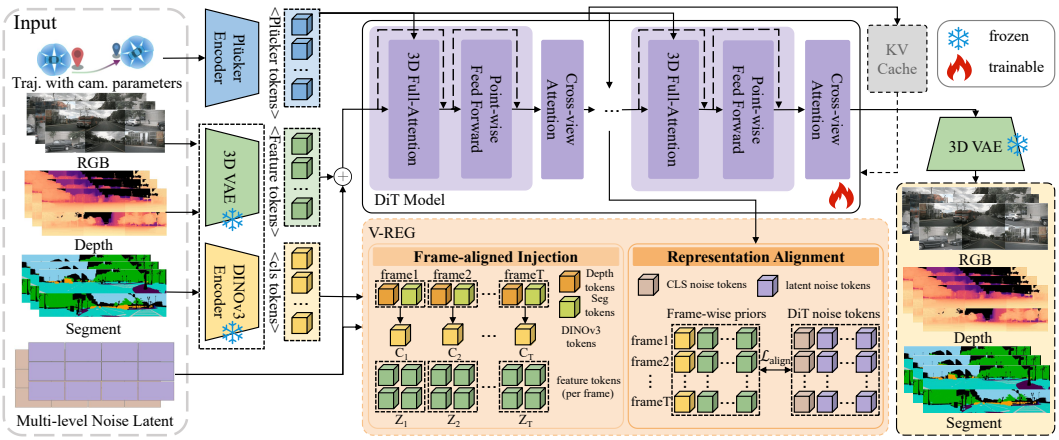


Figure 2: **Architecture of V-REG.** The multimodal latent encoder maps training-time RGB, depth, and semantic targets into a joint noisy latent sequence, while reference observations, camera parameters, and trajectories provide external conditioning. The DiT backbone jointly denoises future RGB, depth, and semantic latents. During training, frozen DINOv3 features provide frame-view CLS conditions and targets for intermediate feature alignment. At inference, DINOv3 retains only its one-shot condition-encoding role, while the alignment branch is disabled.

2.2 Pretrained Visual Representations for Generation

Pretrained visual representations have been widely used to improve diffusion training. Representation-alignment methods match intermediate generative features with features from frozen visual encoders, accelerating convergence and improving generation quality (Yu et al., 2024; Leng et al., 2025; Wang et al., 2026). Other approaches directly inject semantic representations into the denoising process to provide persistent high-level guidance (Wu et al., 2026; Pan et al., 2026). Recent methods such as Align4Gen (Lee et al., 2025) further extend representation alignment to general video generation by preserving correspondence across frames. However, they do not address synchronized multi-camera driving videos, where representations must additionally remain associated with specific camera viewpoints and multimodal scene structures.

3 Methodology

3.1 Overview

Given reference multi-view observations, camera parameters, and an ego trajectory, our model jointly generates future multi-view RGB, depth, and semantic videos. The available depth and semantic conditions are encoded by DINOv3 during both training and inference, while their features additionally serve as alignment targets only during training. We train a multimodal latent diffusion model conditioned on reference observations and geometric trajectory information, where V-REG enhances the DiT-based multi-view video generation framework by incorporating pretrained visual representation guidance. As illustrated in Figure 2, V-REG consists of three components: multimodal condition encoding for constructing structured generation tokens, a DiT-based backbone for latent video denoising, and a frame-aligned REG module combining condition injection with training-only representation alignment.

3.2 Multimodal and Geometric Condition Encoding

The generation process requires both appearance-aware multimodal representations and geometry-aware camera conditions. Therefore, we encode the target modalities and trajectory information into structured tokens that provide complementary guidance for the diffusion backbone.

Multimodal Latent Encoding. We adopt a pretrained 3D VAE (Kong et al., 2024) to encode RGB videos, depth maps, and semantic maps into compact spatio-temporal latent representations. For each modality $m \in \{r, d, s\}$, the corresponding input $\mathbf{x}^m$ is encoded as $\mathbf{z}_0^m = \mathcal{E}_{\text{vae}}(\mathbf{x}^m)$, where

$\mathcal{E}_{\text{vae}}$ denotes the frozen 3D VAE encoder. The modality-specific latents are concatenated along the channel dimension: $\mathbf{z}_0 = \text{Concat}_c(\mathbf{z}_0^r, \mathbf{z}_0^d, \mathbf{z}_0^s)$.

The resulting multimodal latent representation jointly captures appearance, geometric structure, and semantic information, providing a unified representation for multi-modal video generation.

Geometric Trajectory Encoding. To incorporate camera motion and viewpoint constraints, we adopt the normalized Plücker ray-map representation (He et al., 2024). Given camera intrinsics and extrinsics, dense ray maps are constructed and projected into the transformer hidden dimension after spatial-temporal downsampling and patchification: $\mathbf{P} = \rho_{\text{ray}}\left(\text{Patchify}(\text{Down}(\mathbf{R}_{\text{ray}}))\right)$, where $\mathbf{P}$ denotes the resulting geometry-aware Plücker tokens. These tokens are concatenated with the diffusion latent tokens and subsequently processed by the multi-view diffusion backbone.

3.3 Multi-view Diffusion Backbone

Based on the encoded multimodal and geometric tokens, we employ a DiT-based backbone to perform multi-view video denoising. Our backbone is built upon the MMDiT architecture, consisting of 19 dual-stream blocks followed by 38 single-stream blocks. In the dual-stream stage, image tokens $\mathbf{x}$ and trajectory ray-map tokens $\mathbf{r}$ are processed through separate self-attention and MLP pathways with shared modulation parameters derived from the timestep embedding, then interact via joint attention where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are concatenated across both streams. In the single-stream stage, all tokens are merged for unified processing. The denoising process at each step is formulated as:

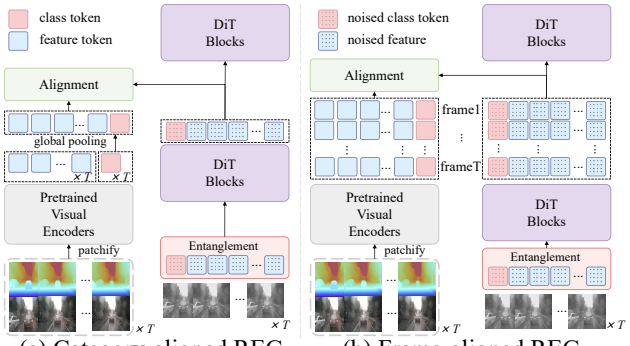
$$\hat{\mathbf{x}}_0 = \frac{\mathbf{x}_t - c_{\text{skip}}(t)\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{r})}{c_{\text{out}}(t)}, \tag{1}$$

where $\boldsymbol{\epsilon}_\theta$ denotes our backbone prediction and c_{skip}, c_{out} are noise-schedule-dependent scaling factors. We adopt multi-level noise where each frame independently samples a noise level, producing per-frame modulation vectors $(\gamma_t, \beta_t, \alpha_t)$ that independently control the affine transformation and gating of each frame's features. For geometric consistency across the 6 surrounding cameras, cross-view attention is applied every 9 single-stream blocks following a ring topology: camera i attends to its two adjacent cameras via dedicated projection layers, efficiently propagating multi-view constraints without the quadratic cost of full pairwise attention. Each block employs RMSNorm-based QK normalization and 3D rotary positional embeddings for temporal, height, and width axes respectively.

3.4 Pretrained Visual Representation Guidance

Although the diffusion backbone receives multimodal and geometric conditions, learning structured representations solely from the denoising objective remains challenging. V-REG therefore uses a frozen DINOv3 encoder through two complementary pathways. Its frame-view CLS features are injected as conditions during both training and inference, while intermediate DiT features are aligned with the same DINOv3 space only during training.

From Category-aligned to Frame-aligned REG. As illustrated in Figure 3, Category-aligned REG globally pools visual representations from all frames into a single video-level class token, which is then entangled with the entire latent sequence and aligned with globally aggregated diffusion features. Although this strategy transfers high-level visual priors, global pooling removes the correspondence between individual frames, viewpoints, and their representations. In contrast, Frame-aligned REG retains a separate representation for each frame-view pair and associates it only with the corresponding diffusion tokens. This design preserves temporal and cross-view correspondence, enabling finer-grained representation injection and



(a) Category-aligned REG (b) Frame-aligned REG

Figure 3: Comparison between Category-aligned and Frame-aligned REG.

alignment. Using one CLS token for each frame-view pair provides a compact trade-off between semantic supervision and computational efficiency. This design preserves temporal and view-specific identity while avoiding the overhead of dense patch-level alignment.

Frame-and-view-aligned Condition Encoding. We use a frozen DINOv3 ViT-B/16 (Siméoni et al., 2025) to encode the available depth and semantic conditions. Although pretrained on natural RGB images, DINOv3 captures transferable boundaries, regions, shapes, and spatial layouts that remain informative in consistently rendered depth and semantic maps. Preprocessing details are provided in Appendix A.1.

For each frame-view pair, DINOv3 extracts depth and semantic CLS tokens: $\mathbf{c}_{t,v} = [\mathbf{c}^d_{t,v}; \mathbf{c}^s_{t,v}]$, which are injected as conditions in both training and inference and detached as alignment targets during training.

Frame-aligned Representation Injection. The depth and semantic CLS tokens are projected into the DiT hidden dimension and added to image tokens belonging to the same frame-view pair. This correspondence-preserving injection supplies each frame with a compact semantic and geometric summary while avoiding dense patch-level conditioning. DINOv3 is evaluated once per input batch, and the resulting CLS tokens are reused throughout iterative denoising. The same injection pathway is applied during training and inference.

Frame-aligned Representation Alignment. To further encourage the diffusion backbone to learn pretrained visual structures, we introduce an intermediate feature alignment objective. Specifically, image hidden states are extracted from layer l of the dual-stream blocks, where the token sequence has shape $B \times (T \cdot V \cdot P) \times d$. Here T, V, and $P{=}H_s{\times}W_s$ denote the number of frames, camera views, and spatial tokens per view, respectively, and d is the hidden dimension.

Spatial tokens within each frame-view pair are mean-pooled: $\bar{\mathbf{z}}_{t,v} = \frac{1}{P}\sum_{i=1}^{P} \mathbf{z}_{t,v,i}$, where $\mathbf{z}_{t,v,i} \in \mathbb{R}^d$ is the i-th spatial token for frame t and view v. The pooled representation is then projected into the DINOv3 embedding space through a lightweight MLP: $\hat{\mathbf{c}}_{t,v} = \psi(\bar{\mathbf{z}}_{t,v}), \quad \psi : \mathbb{R}^d \to \mathbb{R}^{d'}$, where d' is the DINOv3 CLS token dimension.

The alignment loss independently matches each frame-view representation with its corresponding pretrained reference:

$$\mathcal{L}_{\text{align}} = \frac{1}{TV}\sum_{t=1}^{T}\sum_{v=1}^{V}\left(2 - 2\frac{\langle \hat{\mathbf{c}}_{t,v}, \mathbf{c}^{\text{ref}}_{t,v}\rangle}{\|\hat{\mathbf{c}}_{t,v}\|_2 \|\mathbf{c}^{\text{ref}}_{t,v}\|_2}\right). \tag{2}$$

This frame-aligned objective preserves temporal and cross-view structures by avoiding representation collapse caused by global feature aggregation.

The final optimization objective combines the original diffusion loss with the proposed representation alignment loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda \mathcal{L}_{\text{align}}, \tag{3}$$

where λ controls the contribution of pretrained representation guidance. During training, the DINOv3 encoder remains frozen, while the diffusion backbone and projection layers are optimized. Therefore, V-REG improves intermediate representation learning without introducing additional generation branches or modifying the original denoising pathway. During inference, DINOv3 encodes the available depth and semantic conditions once before denoising, and the resulting frame-view CLS tokens are reused at all denoising steps. The alignment projection and loss are not evaluated.

3.5 Spatiotemporal Cache Mechanism

To support efficient chunk-autoregressive generation, we incorporate a dual-component KV cache: a short-term sliding window of the W most recent chunks for local temporal continuity, and a long-term semantic anchor formed by the REG class tokens for globally stable guidance. Since REG introduces only a single class token per chunk, the additional cache overhead is negligible (less than

Method	Multi-view	Video	Multi-modal	Training Steps	GPU	FID↓	FVD↓
DriveGAN (Kim et al., 2021)	✗	✓	✗	210K	N/R	73.40	502.30
BEVControl (Yang et al., 2023)	✓	✗	✗	N/R	N/R	24.85	-
DriveDreamer (Wang et al., 2024b)	✗	✓	✗	N/R	8× A800	14.90	340.80
BEVGen (Swerdlow et al., 2024)	✓	✗	✗	N/R	8× A5000	25.54	-
Vista (Gao et al., 2024b)	✗	✓	✗	N/R	N/R	6.90	89.40
DrivingWorld (Hu et al., 2024)	✗	✓	✗	1000K	64× A100	7.40	90.90
Epona (Zhang et al., 2025)	✗	✓	✗	600K	48× A100	7.50	82.80
DrivingGPT (Chen et al., 2025)	✗	✓	✗	N/R	N/R	12.78	142.61
X-Scene (Yang et al., 2026)	✓	✗	✓	N/R	8× A100	11.29	-
MagicDrive (Gao et al., 2024a)	✓	✓	✗	N/R	8× V100	16.20	-
Panacea (Wen et al., 2024)	✓	✓	✗	N/R	N/R	16.96	139.00
Drive-WM (Wang et al., 2024c)	✓	✓	✗	N/R	8× A40	15.80	122.70
GenAD (Zheng et al., 2024)	✓	✓	✗	N/R	8× RTX 3090	15.40	184.00
DriveDreamer-2 (Zhao et al., 2025)	✓	✓	✗	N/R	8× A800	11.20	55.70
MagicDrive-V2 (Gao et al., 2025)	✓	✓	✗	150K	32× A800	20.91	94.84
WoVoGen (Lu et al., 2024)	✓	✓	✓	N/R	8× A6000	27.60	417.70
OccScene (Li et al., 2025)	✓	✓	✓	N/R	8× A100	11.87	-
UniFuture (Liang et al., 2025)	✓	✓	✓	N/R	8× H20	11.8	99.9
V-REG (Ours)	✓	✓	✓	**30K**	6× H100	8.98	98.75

Table 1: **Quantitative comparison of RGB video generation** on the nuScenes validation set. "Multi-modal" denotes the generation of additional scene-level modalities beyond RGB. Training steps and hardware are taken from the original papers when available; N/R denotes not reported. Since initialization, model scale, resolution, and evaluation protocols differ, cross-paper training costs are contextual rather than strictly compute-normalized.

Method	Input	IoU↑	mIoU↑
AICNet* (Li et al., 2020)	C+D	23.8	10.6
3DSketch* (Chen et al., 2020)	C+D	25.6	10.7
LMSCNet (Roldao et al., 2020)	L	27.3	11.5
JS3C-Net (Yan et al., 2021)	L	30.2	12.5
L-CONet (Wang et al., 2023)	L	30.9	15.8
MonoScene (Cao & De Charette, 2022)	C	18.4	6.9
TPVFormer (Huang et al., 2023)	C	15.3	7.8
C-CONet (Wang et al., 2023)	C	20.1	12.8
HTCL (Li et al., 2024a)	C	21.4	14.1
SparseOcc (Tang et al., 2024)	C	21.8	14.1
OccScene (Li et al., 2025)	C	–	12.2
OccGen (Wang et al., 2024a)	C	23.4	14.5
V-REG (Ours)	C	**31.4**	<u>15.5</u>

Table 2: **Semantic occupancy reconstruction.**

Method	AbsRel↓	δ_1 ↑	δ_2 ↑
UniFuture*(Liang et al., 2025)	0.65	0.28	0.58
V-REG (Ours)	**0.40**	**0.63**	**0.77**

Table 3: **Generated panoramic depth quality.**

ID	Depth	Seg	Align	KVCache	FID↓	FVD↓
1	✗	✗	✗	✗	38.40	152.83
2	✓	✗	✓	✗	35.51	135.99
3	✗	✓	✓	✗	36.95	147.83
4	✓	✓	✗	✗	37.17	140.71
5	✓	✓	✓	✗	29.57	129.79
6	✓	✓	✓	✓	**27.56**	**89.79**

Table 4: **Ablation of Frame-aligned REG.**

0.5% of the cached sequence length), with total memory bounded at $\mathcal{O}(W \cdot L \cdot d)$. As shown in Figure 5(d), the KV cache also mitigates quality degradation over extended future frames compared to the variant without caching.

4 Experiment

4.1 Experimental Setup

Dataset. We conduct experiments on nuScenes (Caesar et al., 2020) using its official scene-level split of 850 training scenes and 150 validation scenes, with video clips sampled within individual scenes to avoid crossing scene boundaries. We further evaluate on PandaSet (Xiao et al., 2021), following the same sequence-level sampling protocol and using separate training and validation splits.

Metrics. For RGB video generation, we report Fréchet Inception Distance (FID) (Heusel et al., 2017) and Fréchet Video Distance (FVD) (Unterthiner et al., 2018). For panoramic depth, we report absolute relative error (Abs. Rel.) and threshold accuracy δ (Long et al., 2021; Zou et al., 2024), using LiDAR-projected depth as ground truth. For downstream 3D semantic occupancy evaluation, the RGB, depth, and semantic videos generated by V-REG are converted into semantic occupancy using an evaluation module implemented following OccScene (Li et al., 2025). We report IoU and

mIoU on the nuScenes-Occupancy validation set. This module is used only for evaluation and is not part of V-REG training.

Multi-view evaluation protocol. For all locally evaluated methods and ablation variants, we use the same validation clips, sequence lengths, spatial resolutions, sampling seeds, and denoising configurations. FID is computed over generated frames from all six camera views using a shared feature extractor. FVD is computed independently for each camera-view video and then averaged across the six views. UniFuture* and all controlled PandaSet variants follow exactly the same evaluation protocol as V-REG. Cross-paper results in Table 1 are taken from the original publications and are therefore treated as contextual rather than strictly protocol-matched comparisons.

Backbone. Our DiT-based model contains 11.23B parameters, including an 11B Diffusion Transformer backbone consistent with prior works and 0.23B newly introduced parameters for cross-view attention layers and the REG module. The backbone is kept largely unchanged, allowing us to study whether pretrained visual representations can improve training efficiency without redesigning the full generation pipeline.

Training details. V-REG is trained on 6 NVIDIA H100 GPUs (80GB) for 30K optimization steps, with a total effective batch size of 6 (1 sample per GPU). We use AdamW with a learning rate of 1×10^{-4} and $\epsilon = 10^{-15}$, gradient clipping at 1.0, and BF16 mixed-precision training with DeepSpeed ZeRO-2 and CPU offloading. The model jointly generates RGB, depth, and semantic videos using variable temporal lengths and spatial resolutions, including 33 frames at 224×400 and 17 frames at 448×800. Frame-aligned REG uses DINOv3-derived depth and semantic CLS tokens, which are computed once per input batch and cached across all denoising steps. We set the alignment loss weight to $\lambda = 0.1$ and zero-initialize the REG projection layers. All reported results are obtained after 30K steps under this training configuration.

4.2 Comparison with State-of-the-Art Methods

RGB Video Generation Quality. As shown in Table 1, V-REG achieves 8.98 FID and 98.75 FVD after only 30K training steps. Compared with existing driving video generators requiring substantially longer training schedules, V-REG reaches competitive visual quality and temporal consistency with significantly fewer optimization iterations, demonstrating the effectiveness of pretrained visual representations for accelerating multi-view video generation. Since cross-paper configurations differ, Table 1 is intended as a contextual comparison rather than a strictly compute-normalized benchmark. The controlled PandaSet study in Figure 5 isolates the effect of representation guidance under identical settings.

Generated Panoramic Depth Quality. For a controlled comparison, we re-evaluate UniFuture using the same depth prior as V-REG, denoted as UniFuture* in Table 3. V-REG reduces Abs. Rel. from 0.65 to 0.40 and improves $\delta < 1.25$ and $\delta < 1.25^2$ from 0.28 and 0.58 to 0.63 and 0.77, respectively. As shown in Figure 4, V-REG also produces clearer object boundaries and finer local geometry, indicating that its gains arise from improved multi-modal representation learning rather than the depth prior alone.

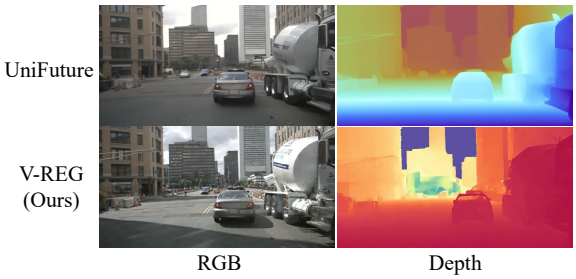
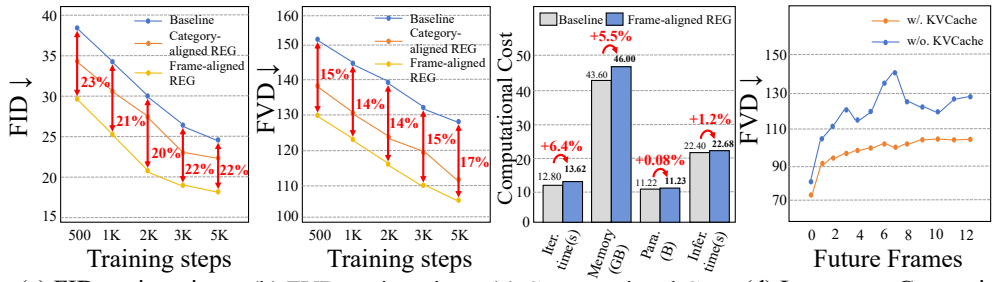


Figure 4: Qualitative comparison of generated RGB images and depth maps.

Generated Semantic Occupancy Quality. We evaluate semantic occupancy generation using IoU and mIoU. AICNet*(Li et al., 2020) and 3DSketch*(Chen et al., 2020) utilize both camera-based RGB images and LiDAR-projected depth maps as inputs, following OpenOccupancy(Wang et al., 2023). As shown in Table 2, V-REG achieves state-of-the-art IoU and the second-best mIoU among the compared methods. Notably, V-REG obtains competitive results even compared with LiDAR-based approaches (Roldao et al., 2020; Wang et al., 2023), which benefit from precise 3D geometric measurements, while relying on camera-based inputs rather than LiDAR measurements. These results indicate that V-REG can generate not only visually plausible videos, but also structured semantic scene representations.



(a) FID vs. iterations (b) FVD vs. iterations (c) Computational Cost (d) Long-term Generation

Figure 5: **Controlled convergence and computational cost.** (a,b) FID and FVD from 500 to 5K iterations under identical settings. Red annotations show the relative reductions of Frame-aligned REG over the baseline. (c) Normalized computational costs, with absolute values shown above the bars. (d) FVD over increasing future frames with and without KV Cache.

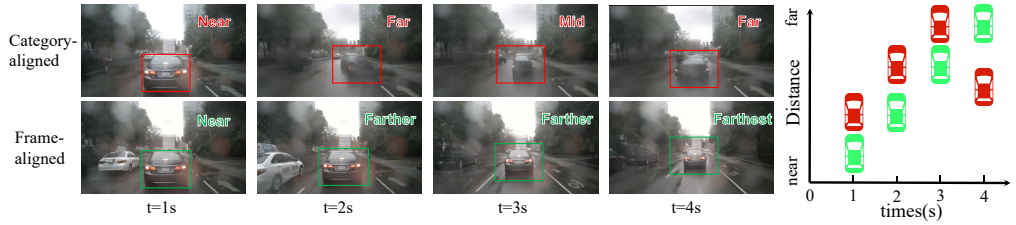


Figure 6: **Temporal-spatial consistency of representation guidance.** Category-aligned REG produces a non-monotonic Near–Far–Mid–Far progression for the tracked vehicle, whereas Frame-aligned REG preserves a more coherent Near–Farther–Farther–Farthest evolution from $t = 1$ s to $t = 4$ s.

Overall, the RGB, depth, and occupancy results show that V-REG improves the efficiency-quality trade-off of multi-view driving video generation while maintaining the ability to model structured scene geometry and semantics.

4.3 Ablation Study

Unless otherwise specified, all variants are initialized from the same pretrained checkpoint and trained on PandaSet using 4 NVIDIA RTX 5880 Ada Generation GPUs under identical optimization settings.

Controlled Convergence and Efficiency. Figure 5(a,b) compares the baseline, Category-aligned REG, and Frame-aligned REG on PandaSet under identical settings. Frame-aligned REG consistently achieves the lowest FID and FVD from 500 to 5K iterations, with Category-aligned REG providing intermediate gains, confirming that pretrained representations accelerate convergence while frame-wise correspondence further improves fidelity. Figure 5(c) validates on nuScenes that these gains introduce only modest overhead. Figure 5(d) further shows that the proposed KV Cache effectively preserves FVD over extended temporal horizons on PandaSet, demonstrating its necessity for long-horizon autoregressive video synthesis.

Qualitative Spatiotemporal Consistency. Figure 6 further illustrates the effect of frame-wise alignment by tracking the highlighted foreground vehicle under forward ego motion. As the ego vehicle moves forward, the relative distance to the foreground vehicle should increase progressively. Category-aligned REG instead produces a non-monotonic *Near–Far–Mid–Far* progression, revealing inconsistent object scale and spatial evolution across frames. In contrast, Frame-aligned REG generates a more coherent *Near–Farther–Farther–Farthest* sequence from $t = 1$ s to $t = 4$ s. This qualitative result is consistent with the FVD convergence in Figure 5(b), further confirming improved spatiotemporal consistency.

Effect of Depth and Semantic Guidance. Table 4 evaluates depth guidance, semantic guidance, representation alignment, and KV cache. Depth guidance alone achieves 35.51 FID and 135.99 FVD (row 2), while semantic guidance alone yields 36.95 FID and 147.83 FVD (row 3). Both improve upon the baseline (38.40/152.83), with depth features providing stronger gains by directly encoding scene geometry and spatial layout. Combining depth and semantic representations without explicit alignment yields 37.17 FID and 140.71 FVD (row 4), indicating that direct feature injection alone cannot fully exploit pretrained visual priors. The full model with alignment achieves 29.57 FID and 129.79 FVD (row 5), demonstrating the effectiveness of multimodal guidance and frame-

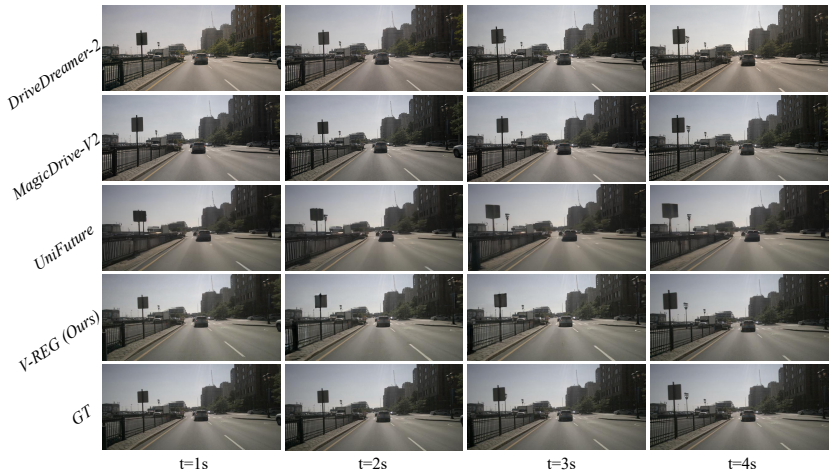


Figure 7: Qualitative comparison with existing driving video generation methods. Compared with MagicDrive-V2, DriveDreamer-2, and UniFuture, V-REG better preserves scene geometry, object layouts, local details, and temporal consistency.

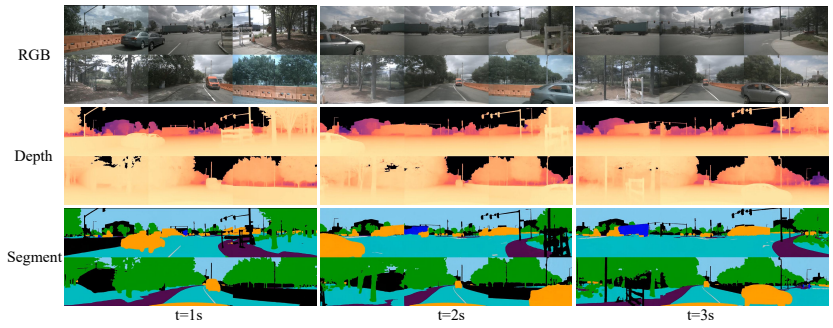


Figure 8: Qualitative visualization of V-REG outputs. We show generated RGB videos, panoramic depth maps, and semantic maps from $t = 1\,\text{s}$ to $t = 3\,\text{s}$. V-REG produces temporally coherent multi-modal driving scenes, with depth and semantic predictions structurally aligned with the generated RGB content.

wise alignment. Furthermore, incorporating KV cache for inference acceleration further improves performance to 27.56 FID and 89.79 FVD (row 6), as cached attention states enable more consistent temporal modeling across frames.

4.4 Qualitative Multimodal Generation Results

As shown in Figure 7, existing methods progressively lose structural detail, exhibiting blurred lane boundaries and distorted vehicles by $t = 4\,\text{s}$, whereas V-REG better preserves road layouts, object geometry, and temporal evolution. Figure 8 further shows that the jointly generated depth and semantic maps remain spatially and temporally aligned with the RGB video, particularly along object boundaries and drivable regions. These observations complement the quantitative results in Tables 1, 3, and 2, supporting improved visual quality, temporal consistency, and cross-modal coherence.

5 Conclusion

We propose V-REG, a visual representation-guided framework for training-efficient multi-view driving video generation. By introducing DINOv3-based semantic priors into a DiT-based diffusion backbone, V-REG reduces the burden of learning complex driving-scene semantics, while its frame-aligned representation entanglement and alignment strategy preserves temporal and multi-view consistency. Experiments show that V-REG achieves competitive RGB video generation quality with only 30K training steps, together with strong panoramic depth generation and downstream semantic occupancy reconstruction performance. These results validate the effectiveness of pretrained visual guidance in improving the efficiency–quality trade-off, while future work will explore longer-horizon generation and closed-loop evaluation.

AI USE STATEMENT

In this work, we used generative AI tools to edit the manuscript and improve the readability of the English text. We have reviewed all AI-assisted text and verified that the technical content, claims, and reported numbers originate from our own experiments and derivations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on improving the training efficiency and generation quality of multi-view driving video generation models. All experiments are conducted on the publicly available nuScenes dataset (Caesar et al., 2020) and PandaSet dataset (Xiao et al., 2021), which are used for their intended research purposes in autonomous driving. No new human data was collected for this study. The proposed method (V-REG) leverages a frozen pretrained DINOv3 encoder (Siméoni et al., 2025) to provide visual representation guidance during training, which is removed at inference time and does not introduce additional ethical concerns beyond those inherent to large-scale video generation. We acknowledge that, like all generative models, our method could potentially be misused to produce misleading or harmful driving scenarios, and we encourage responsible use of the technology.

REPRODUCIBILITY STATEMENT

All experiments are conducted on the public nuScenes and PandaSet benchmarks; the dataset splits, evaluation metrics, and evaluation protocols are described in Sec. 4.1. The nuScenes dataset uses the official scene-level split of 850 training scenes and 150 validation scenes, while PandaSet follows the same sequence-level sampling protocol with separate training and validation splits. Complete architecture details and training hyperparameters are provided in Sec. 4.1.5, including the use of NVIDIA GPUs, AdamW optimizer with learning rate 1×10^{-4}, BF16 mixed-precision training with DeepSpeed ZeRO-2, and training for 30K optimization steps. The comparison with prior methods follows consistent evaluation protocols where applicable, with FID and FVD computed over generated frames from all six camera views. Detailed ablation studies and additional implementation details are provided in the supplementary material. We will release the source code upon publication.

REFERENCES

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*, pp. 11621–11631, 2020.

Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3991–4001, 2022.

Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4193–4202, 2020.

Yuntao Chen, Yuqi Wang, and Zhaoxiang Zhang. Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26890–26900, 2025.

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*, volume 2024, pp. 22841–22860, 2024a.

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*, pp. 28135–28144, 2025.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024b.

Xingtai Gui, Meijie Zhang, Tianyi Yan, Wencheng Han, Jiahao Gong, Feiyang Tan, Cheng-zhong Xu, and Jianbing Shen. Bridging scene generation and planning: Driving with world model via unifying vision and motion representation. *arXiv preprint arXiv:2603.14948*, 2026.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. Drivingworld: Constructing world model for autonomous driving via video gpt. *arXiv preprint arXiv:2412.19505*, 2024.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9223–9232, 2023.

Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5820–5829, 2021.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Dohun Lee, Hyeonho Jeong, Jiwook Kim, Duygu Ceylan, and Jong Chul Ye. Improving video diffusion transformer training by multi-feature fusion and alignment from self-supervised vision encoders. *arXiv preprint arXiv:2509.09547*, 2025.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18262–18272, 2025.

Bohan Li, Jiajun Deng, Wenyao Zhang, Zhujin Liang, Dalong Du, Xin Jin, and Wenjun Zeng. Hierarchical temporal context learning for camera-based semantic scene completion. In *European Conference on Computer Vision*, pp. 131–148. Springer, 2024a.

Bohan Li, Xin Jin, Jianan Wang, Yukai Shi, Yasheng Sun, Xiaofeng Wang, Zhuang Ma, Baao Xie, Chao Ma, Xiaokang Yang, et al. Occscene: Semantic occupancy-based cross-task mutual learning for 3d scene generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3351–3359, 2020.

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

Yaoru Li, Federico Landi, Marco Godi, Xin Jin, Ruiju Fu, Yufei Ma, Muyang Sun, Heyu Si, and Qi Guo. Far-drive: Frame-autoregressive video generation in closed-loop autonomous driving. *arXiv preprint arXiv:2603.14938*, 2026.

Dingkang Liang, Dingyuan Zhang, Xin Zhou, Sifan Tu, Tianrui Feng, Xiaofan Li, Yumeng Zhang, Mingyang Du, Xiao Tan, and Xiang Bai. Seeing the future, perceiving the future: A unified driving world model for future generation and perception. *arXiv preprint arXiv:2503.13587*, 2025.

Xiaoxiao Long, Lingjie Liu, Wei Li, Christian Theobalt, and Wenping Wang. Multi-view depth estimation using epipolar spatio-temporal networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8258–8267, 2021.

Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *European conference on computer vision*, pp. 329–345. Springer, 2024.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

Yueming Pan, Ruoyu Feng, Qi Dai, Yuqi Wang, Wenfeng Lin, Mingyu Guo, Chong Luo, and Nanning Zheng. Semantics lead the way: Harmonizing semantic and texture modeling with asynchronous latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43664–43674, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pp. 111–119. IEEE, 2020.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters*, 9(4):3578–3585, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15035–15044, 2024.

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

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. Occgen: Generative multi-modal 3d occupancy prediction for autonomous driving. In *European Conference on Computer Vision*, pp. 95–112. Springer, 2024a.

Mengmeng Wang, Dengyang Jiang, Liuzhuozheng Li, Yucheng Lin, Guojiang Shen, Xiangjie Kong, Yong Liu, Guang Dai, and Jingdong Wang. Sra 2: Variational autoencoder self-representation alignment for efficient diffusion training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32978–32987, 2026.

Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17850–17859, 2023.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pp. 55–72. Springer, 2024b.

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

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *Conference on robot learning*, pp. 539–549. PMLR, 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*, pp. 6902–6912, 2024.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Ming-Ming Cheng, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. *Advances in Neural Information Processing Systems*, 38:7714–7743, 2026.

Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE international intelligent transportation systems conference (ITSC)*, pp. 3095–3101. IEEE, 2021.

Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 3101–3109, 2021.

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.

Yu Yang, Alan Liang, Jianbiao Mei, Yukai Ma, Yong Liu, and Gim Hee Lee. X-scene: Large-scale driving scene generation with high fidelity and flexible controllability. *Advances in Neural Information Processing Systems*, 38:104415–104451, 2026.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. Epona: Autoregressive diffusion world model for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27220–27230, 2025.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10412–10420, 2025.

Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*, pp. 87–104. Springer, 2024.

Yingshuang Zou, Yikang Ding, Xi Qiu, Haoqian Wang, and Haotian Zhang. M 2 depth: Self-supervised two-frame m ulti-camera m etric depth estimation. In *European Conference on Computer Vision*, pp. 269–285. Springer, 2024.

A Additional Implementation Details

This supplementary material provides additional implementation details and qualitative results that complement the main paper. In particular, we provide further details on the preprocessing of non-RGB modalities, additional comparisons with existing driving video generation methods, and qualitative visualizations of the jointly generated RGB, depth, and semantic sequences.

A.1 Preprocessing of Non-RGB Representation Targets

V-REG employs a frozen DINOv3 encoder to extract pretrained visual representations from the depth and semantic targets available during training. Although DINOv3 is pretrained primarily on natural RGB images, consistently rendered depth and semantic maps preserve strong region, boundary, shape, and spatial-layout structures. These structures remain informative when processed by a self-supervised visual encoder and provide complementary geometric and semantic guidance to the diffusion backbone.

For depth preprocessing, metric depth values are clipped to the range $[0, 100]$ meters. Sky pixels, identified from the corresponding semantic labels, are assigned the maximum depth value. The resulting single-channel depth map is replicated to three channels and linearly rescaled before being passed to the pretrained visual encoder.

For semantic preprocessing, the semantic labels are rendered using a fixed nine-class RGB palette. Depth and semantic maps undergo the same spatial transformations as their corresponding RGB frames to preserve frame-view correspondence across modalities. Before DINOv3 encoding, both modalities are resized to 224×224, rescaled to the input range expected by the pretrained encoder, and normalized using the standard DINOv3 preprocessing configuration.

This unified preprocessing avoids introducing separate modality-specific representation encoders. More importantly, it preserves the temporal and camera-view correspondence required by Frame-aligned REG while exposing shared boundary, region, and shape-level structures to the frozen visual encoder. We use DINOv3 ViT-B/16 with an embedding dimension of 768 throughout all experiments.

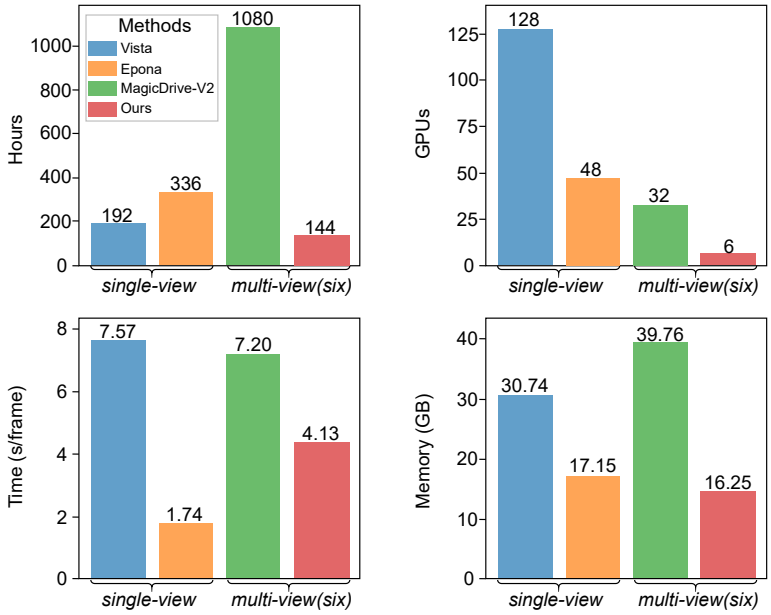


Figure 9: **Comparison of Video Generation Efficiency.** We compare our method with previous approaches in terms of training time, number of GPUs used for training, inference time, and VRAM consumption during inference.

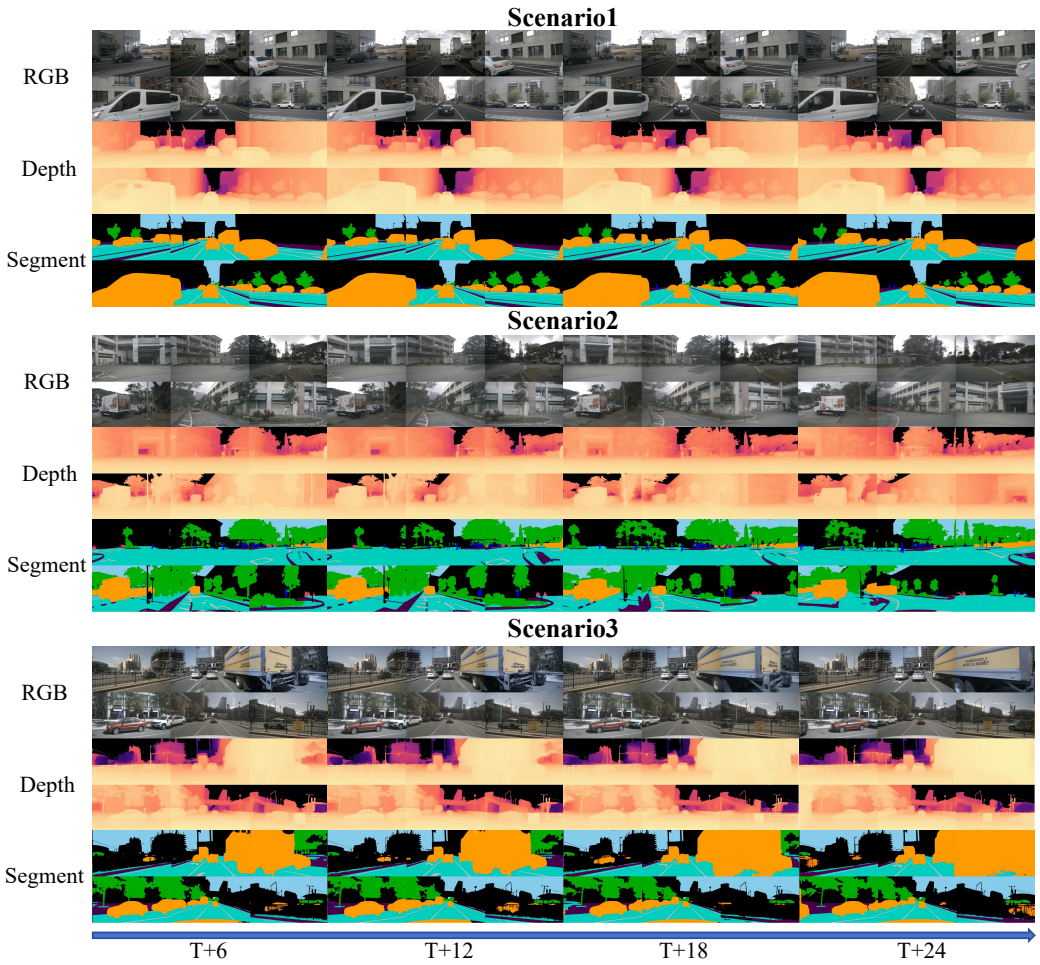


Figure 10: More multi-modal generation results, including RGB videos, panoramic depth maps, and semantic maps.

B Efficiency of Video Generation

Previous video-based driving world models or driving video generation models are typically trained from scratch, which is computationally expensive. We summarize the training and inference efficiency of different methods in Figure 9. As shown in the figure, previous works such as Vista (Gao et al., 2024b), Epona (Zhang et al., 2025), and MagicDrive-V2 (Gao et al., 2025) require a large number of GPUs (32–128) and long training periods (192–1080 h). In contrast, we leverage pre-trained video generation models to reduce training cost. With our twostage generation framework and fine-tuning strategy, the total training time is reduced to less than a week (3 days for training the first two stages and 3 days for the final stage) using only 6 GPUs. Moreover, our method also achieves superior inference efficiency: the average generation speed reaches 4.13 s per frame, and the VRAM consumption is only 16.25 GB. MagicDrive-V2 requires 39.76 GB of VRAM by performing sequence parallelism across 8 GPUs, whereas other methods, including ours, can perform inference on a single GPU. Vista and Epona generate only single-view videos, while MagicDrive-V2 and our method support six-view video generation.

C More Video Generation Results

Figure 10 visualizes the RGB, panoramic depth, and semantic sequences jointly generated by V-REG. We show multiple consecutive timesteps to examine whether the generated modalities remain spatially and temporally consistent with one another.

The generated panoramic depth maps exhibit scene structures that closely correspond to the RGB content. Foreground vehicles, road surfaces, and distant scene elements appear at consistent spatial locations across the two modalities. Similarly, the semantic predictions preserve the major object and road regions observed in the RGB sequence.

Importantly, this correspondence is maintained over time rather than only within individual frames. As objects move relative to the ego vehicle, their locations and structural boundaries evolve consistently across RGB, depth, and semantic predictions. This indicates that V-REG does not simply improve RGB appearance, but also enables the multimodal diffusion backbone to maintain a coherent shared scene representation across modalities.

Together with the panoramic depth and downstream semantic occupancy results reported in the main paper, these qualitative examples demonstrate that the generated sequences retain useful geometric and semantic structures beyond visual appearance.