This supplementary material provides additional details to support the findings presented in our main paper. We include: (1) comprehensive implementation specifics for our RLGF framework and the pre-trained perception models; further details on the GeoScores metric computation; (2) additional experimental results (3) a discussion on the limitations of our current work and potential future directions.

A Detailed Implementation Details

This section elaborates on the implementation details of our proposed RLGF framework, the pre-trained perception models ($\mathcal{P}_{geo}$ and $\mathcal{P}_{occ}$) , and the experimental setup.

A.1 Dataset Preparation

All experiments, including the pre-training of our perception models ($\mathcal{P}_{geo}$ and $\mathcal{P}_{occ}$) and the fine-tuning of diffusion models with RLGF, are conducted using the nuScenes dataset [5]. We primarily utilize the official training and validation splits. While nuScenes provides rich annotations like 3D bounding boxes and HD maps, it does not directly offer ground truth labels for vanishing points (VP), dense segmentation masks for all relevant classes (like fine-grained lanes beyond HD map polylines), or per-pixel depth maps required by our $\mathcal{P}_{geo}$. Therefore, we generate high-quality pseudo-labels for these tasks using strong, pre-existing perception models, as detailed below. These pseudo-labels serve as the training targets for our latent-space perception models.

Depth Pseudo-Labels: To obtain dense depth information for training the depth estimation component of $\mathcal{P}_{geo}$, we utilize Depth Anything V2 (vit-l version) [52]. This state-of-the-art monocular depth estimation model is applied to all images in the nuScenes training set to generate per-pixel depth maps. These output depth maps serve as the pseudo-ground truth for our latent depth estimation task.

Semantic Segmentation Pseudo-Labels (Lanes, Road Surface, Vehicles): For precise segmentation masks of various scene elements, we employ Grounded-SAM-2[34, 33]. For the lanes, the model is prompted to accurately segment visible lane markings. The resulting binary segmentation masks are used as pseudo-ground truth for training the lane parsing head of $\mathcal{P}_{geo}$. For the road surface and vehicle masks, SAM-2 is also utilized to generate segmentation masks for road surfaces and vehicles.

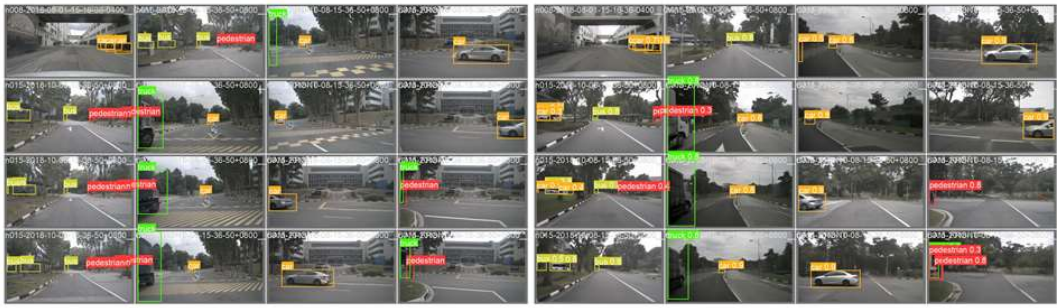
Vanishing Point Pseudo-Labels from Lane Masks, following [11]: With accurate lane segmentation masks obtained via SAM-2 (as described above), we derive vanishing point pseudo-labels through a geometric procedure. For each detected lane marking in a frame: The center point of the lane marking is calculated from its left and right edges (derived from the SAM-2 segmentation mask) for every horizontal image line at 5-pixel intervals. These extracted center points are grouped to represent the centerline of each individual lane marking. Robust curve fitting (e.g., RANSAC with a line model) is applied to these centerlines. The intersection point of multiple fitted lane centerlines is then computed to determine the scene's vanishing point. This computed VP serves as the pseudo-ground truth for the VP detection task.

The use of these high-quality pseudo-labels enables us to train effective latent-space perception models tailored to the nuScenes domain, which subsequently provide the nuanced reward signals for our RLGF framework. The conditions c for the main diffusion models (e.g., semantic 3D boxes for some baselines) are derived from the original nuScenes ground truth annotations.

A.2 Perception Model Architectures and Pre-training

Micro-Decode Module($\mathcal{F}_{micro}$): The $\mathcal{F}_{micro}$ module is constructed using the first upper block of the official VAE decoder from the Latent Diffusion Model [35, 59] used by our baseline video diffusion models. The input of $\mathcal{F}_{micro}$ is the noisy latent feature z_k^f for a frame f with the timestep k. The same $\mathcal{F}_{micro}$ architecture is used when processing the reference real-video latent z_v (with k typically set to 0).

Latent Geometry Perception Model($\mathcal{P}_{geo}$): We use a pre-trained DINOv2-ViT-S/14 [29] as the backbone feature extractor and the pre-trained weight from DepthAnything-V2 [52]. $\mathcal{P}_{geo}$ is trained for 50 epochs on the nuScenes training split using the AdamW optimizer with a learning rate of 5×10^{-5} and a batchsize of 16. We use 8$\times$ NVIDIA A100 GPUs to cover the experiment.



Real Video Synthetic Video from DiVE

Figure 4: **Left:** Detection results on a real nuScenes image. **Right:** Detection results on a corresponding synthetic image generated by the DiVE baseline. Bounding boxes indicate detected objects (primarily vehicles).

Latent Occupancy Prediction Model($\mathcal{P}_{occ}$): $\mathcal{P}_{occ}$ is trained on occ3D-nuscenes dataset for 24 epochs using AdamW optimizer following [55, 14].

A.3 RLGF Fine-tuning Details

Baselines: We used publicly available checkpoints for MagicDrive-V2 [7] and DiVE [16].

LoRA Configuration: For LoRA, we applied it to the attention layers (Q, K, V projections) of the DiT backbone in the diffusion models. We used a rank $r = 16$ following [1].

Latent-Space Windowed Optimization: The window size w is set to 5. The starting step t' for the window was randomly sampled from the range $[8, 30]$, with $T = 30$ is the total number of diffusion steps.

Reward Weights: We set $\lambda_{vp} = 0.1, \lambda_{lane} = 0.1, \lambda_{depth} = 0.5$. These weights were determined empirically based on early experiments on a small validation subset, aiming to balance the scale of individual reward components and their perceived impact on generation quality.

We use AdamW with a learning rate of 1×10^{-4} and a batchsize of 1 with 8 frames per video clip.

A.4 GeoScores Metric Details

This section provides further clarification on the computation of our GeoScores components. For all GeoScores, the "reference ground truth" is derived by applying the corresponding pre-trained perception model to the *real* video data, following appendix A.1. The score then measures the deviation of the synthetic video's perception output from this real-data-derived reference.

Vanishing Point Error (VP↓): Calculated as the L2 Normalized Distance (NormDist) between the calculated VP on a synthetic frame and the VP calucalted on a real frame. **Lane Topology Score (Lane↑):** Calculated as the F1-score of the semantic segmentation of lane markings. The predictions is from Grounded-SAM2 [34, 33] on the synthetic frame, and the target is applied to the real frame. **Depth Error (Depth)↓:** Calculated as the Root Mean Squared Error (RMSE) between the depth map predicted by Depth Anything V2 [52] for road surface regions on a synthetic frame and the depth map for the same regions on the real frame. Road surface masks are obtained from SAM-2 [33].

B Additional Experiment Results

B.1 2D Object Detection Results

To illustrate that current diffusion models like DiVE can generate visually realistic data with minimal 2D domain gap for certain tasks, we present qualitative 2D object detection results. Figure 4 shows outputs from a YOLOv5 [17] detector applied to (a) real nuScenes data and (b) synthetic data generated by the DiVE baseline. The detector is pre-trained on a large-scale dataset (e.g., COCO) and then fine-tuned on real nuScenes training data.

As observed in Figure 4, the 2D detection performance on DiVE-generated synthetic data is qualitatively very similar to that on real data. Objects are generally detected with comparable confidence and bounding box accuracy. This visual consistency aligns with our quantitative findings (mAP: 43.8 on synthetic vs. 44.7 on real, as mentioned in the Introduction), suggesting that the semantic content and 2D appearance features necessary for 2D detection are well-preserved in the synthetic videos. This further reinforces our hypothesis that the primary limitation of such synthetic data lies in its 3D geometric fidelity, which is specifically addressed by our RLGF framework.

B.2 Extended 3D Object Detection Results on Multiple Detecters

To further demonstrate the generalizability of the improvements conferred by RLGF, we evaluated the generated synthetic data using an additional state-of-the-art 3D object detector, StreamPETR [43], alongside the BEVFusion results presented in the main paper. Table 5 presents the performance (mAP and NDS on nuScenes validation) for StreamPETR and the average performance across both BEVFusion and StreamPETR. Both detectors were trained from scratch solely on the respective synthetic data or real data.

Table 5: Detailed 3D Object Detection (3DOD) performance on nuScenes validation using StreamPETR [43]. RLGF is applied to MagicDrive-v2 and DiVE.

Methods	Quality	BevFusion		StreamPETR	
	FVD	mAP	NDS	mAP	NDS
Real Data	-	35.53	41.20	38.01	49.02
Panacea [46]	139.0	11.58	22.31	-	-
Drive-WM [46]	122.7	20.66	-	-	-
MagicDrive-v2 [16]	101.2	18.95	21.10	22.77	28.93
DiVE [16]	68.4	25.75	33.61	29.19	36.23
MagicDrive-v2+Ours	99.8	23.21	27.80	26.01	35.64
DiVE+Ours	**67.6**	**31.42**	**36.07**	**33.94**	**39.68**

C Limitations and Future Work

This section discusses the current limitations of our RLGF framework and GeoScores metric, alongside potential avenues for future research.

Dependence on Perception Models: RLGF's performance is inherently tied to the accuracy and robustness of the pre-trained perception models ($\mathcal{P}_{geo}, \mathcal{P}_{occ}$). Biases or errors in these models could propagate into the reward signal and mislead the generation process. Future work could explore jointly training or adapting perception models during RLGF, or using ensembles.

Computational Cost: While Latent-Space Windowed Optimization significantly reduces costs compared to full rollouts, RL-based fine-tuning remains more computationally intensive than standard diffusion model training. Exploring more sample-efficient RL algorithms or distillation techniques could be beneficial.

Reward Design and Balancing: The current HGA reward combines five components with manually tuned weights. Optimizing these weights automatically or learning a more adaptive reward function is a promising direction. Furthermore, incorporating even more diverse geometric or physical constraints (e.g., collision avoidance, traffic rule adherence) could further enhance realism.

Generalization: While demonstrated on nuScenes, further investigation is needed to assess RLGF's generalization capabilities across diverse datasets, environmental conditions (e.g., adverse weather, night scenes not well-represented in training), and different diffusion model architectures.

GeoScores Scope: Current GeoScores focus on camera-based geometric aspects. Expanding them to include LiDAR consistency or multi-modal geometric agreement could provide a more holistic evaluation.