PerLDiff: Controllable Street View Synthesis Using Perspective-Layout Diffusion Model

Supplementary Material

The supplementary material is organized into the following sections:

- Section A: DDPM Preliminaries
- Section B: Implementation Details
- Section C: Limitation and Future Work
- Section D: Additional Experiments
- Section E: Visualization Results

A. DDPM Preliminaries

Denoising Diffusion Probabilistic Models (DDPM) [3] are a class of generation models which simulate a Markov chain of diffusion steps to gradually convert data samples into pure noise. The generative process is then reversed to synthesize new samples from random noise. We commence with an observation x_0 sampled from the data's true distribution $q(x)$, and then progressively apply Gaussian noise over a series of T time steps. The forward diffusion is mathematically defined as $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}),$, where β_t is a variance term that can be either time-dependent or learned during training. The entire forward diffusion process can be represented as the product of the conditional distributions from each step:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{1}$$

where the sequence $\{\beta_t\}_{t=1}^{T}$ specifies the noise schedule applied at each timestep. The diffusion process is notable for permitting direct sampling of x_t from x_0 using a closed-form expression:

$$q(x_t|x_0) = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{2}$$

in which $\alpha_t = 1 - \beta_t$ and the cumulative product $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. To synthesize new samples, a reverse process known as the backward diffusion is learned, which conceptually undoes the forward diffusion. This inverse transition is captured through a parameterized Gaussian distribution:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t), \sigma_\theta^2(x_t)\mathbf{I}). \tag{3}$$

B. Implementation Details

PerLDiff utilizes the pre-trained Stable Diffusion v1.4 [9], augmented with specific modifications to enhance scene control. Training was conducted on a server equipped with eight Tesla V100 (32 GB) GPUs over 60,000 iterations, which required two days. An initial batch size of 16 was adjusted to a per-GPU batch of two for focused optimization, particularly for data samples comprising six view images per frame. The generation of samples conforms to the CFG rule [2], employing a guidance scale of 5.0 and the DDIM [10] across 50 steps.

For scene manipulation, the text encoder within Stable Diffusion is retained, along with a weight-frozen CLIP to manage textual inputs and ConvNext for processing road maps. Feature extraction from PerL boxes is conducted via an MLP, optimized through PerL-based controlling module (PerL-CM) with randomly initialized weights. In contrast, certain modules inherit and freeze pre-trained weights from Stable Diffusion. The key parameters within PerL-CM, λ_b and λ_s, are set to 5.0 to facilitate optimal image synthesis. Furthermore, DDIM [10] and CFG [2] are integrated into our training regimen, with a novel approach of omitting all conditions at a rate 10% to foster model versatility.

The optimization process employs AdamW [8] without a weight decay coefficient and with a learning rate of 5×10^{-5}, complemented by a warm-up strategy during the first 1,000 iterations. BEVFormer [6], StreamPETR [11], and CVT [16] were retrained using original configurations tailored to our target resolution. The performance of BEVFusion [7] and MonoFlex [15] was assessed using their provided code and pre-trained weights.

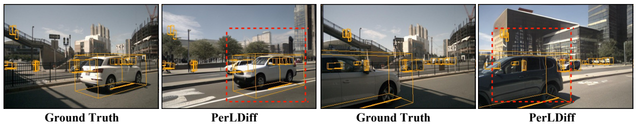


Figure 1. Failure cases of PerLDiff, with red markers highlighting instances where, compared to the ground truth, PerLDiff generates images with the front and rear of vehicles reversed.

C. Limitation and Future Work.

Fig. 1 depicts several failure cases of PerLDiff, where the model erroneously generates vehicles with the front and rear orientations reversed, in contrast to the ground truth. This limitation arises from the usage of a PerL mask in PerLDiff, which does not account for the orientation on the 2D PerL plane. Future endeavors may explore video generation, extending to work such as DrivingDiffusion [4], Panacea [13], and Driving into the Future [12].

Table 1. Controllability comparison for street view image generation on the NuScenes *validation* set. A quantitative evaluation using 3D object detection metrics from BEVFusion [7].

Method	FID↓	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓
Oracle	–	35.54	41.20	0.67	0.27	0.56
MagicDrive [1]	16.59	20.85	30.26	–	–	–
BEVControl*	15.94	13.19	19.91	0.94	0.34	0.96
PerLDiff (Ours)	**15.67**	**24.69**	**30.71**	**0.82**	**0.28**	**0.76**

D. Additional Experiments

In this section, we present additional experiments conducted to validate controllability at different resolutions (256× 704) and to assess the contributions of individual components within our PerLDiff. Our studies focus on the following aspects:

- Effectiveness of Controllable Generation on NuScenes (Subsection D.1)
- Effectiveness of Perl-based Cross Attention (Object) (Subsection D.2)
- Effectiveness of View Cross-attention for Multi-View Consistency (Subsection D.3)
- Effectiveness of PerLDiff Based on ControlNet (Subsection D.4)
- Effectiveness of Classifier-Free Guidance Scale (Subsection D.5)

Our results confirm the superior performance of our method across various resolutions and illustrate how each component is integral to the success of our PerLDiff.

D.1. Effectiveness of Controllable Generation on NuScenes

In Tab. 1, we conduct a comparative analysis to emphasize the capabilities of PerLDiff for controllable generation at a resolution of 256×704. This quantitative evaluation contrasts our method with other leading approaches based on the detection metrics provided by BEVFusion [7]. Our PerLDiff exhibits significantly superior performance, achieving mAP improvements of 3.84% and 11.50%, and NDS increases of 0.45% and 10.80%, compared to MagicDrive [1] and BEVControl*, respectively. These results confirm the efficacy of PerLDiff in the precise controllable generation at the object level.

D.2. Effectiveness of Perl-based Cross Attention (Object)

To facilitate a better understanding of PerLDiff, we provide a detailed explanation of the PerL-based cross-attention (Object). As shown in Fig. 2, MagicDrive [1] utilizes text cross-attention from Stable Diffusion to implicitly learn a unified feature that concatenates text, camera parameters, and bounding boxes in the token dimension. In contrast, PerLDiff employs the PerL masking map as a prior, allowing each object condition to precisely control the corresponding pixel features. This results in more accurate positioning and orientation of objects in the generated images. Additionally, we integrated the object mask into the token dimension corresponding to the bounding box. As shown in Tab. 2, the results indicate improvements in BEVFormer, with NDS (e.g., 29.77 vs. 28.79 for MagicDrive) and mAOE (e.g., 0.73 vs. 0.81 for MagicDrive) demonstrating the effectiveness of PerLDiff in enhancing the performance of MagicDrive. Note that MagicDrive utilizes a single attention map for managing text, camera parameters, and boxes in the cross-attention process. Consequently, our ability to make improvements is constrained by the limited scope available for modifying the attention map within this architecture.

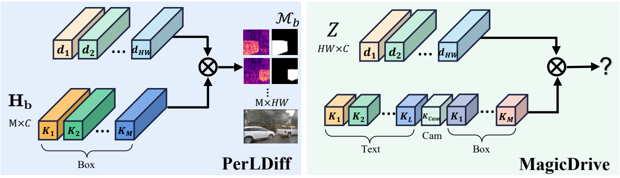


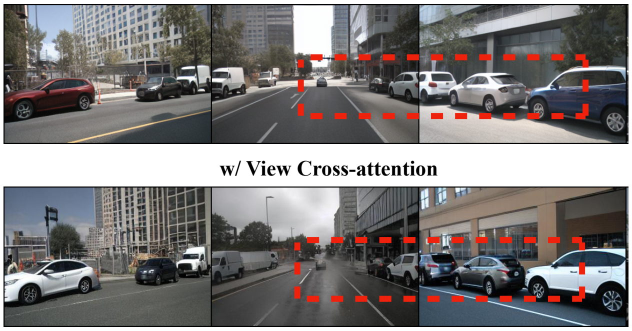
Figure 2. Overview of the PerL-based cross-attention (Object). MagicDrive employs text cross-attention to create a unified feature, while PerLDiff uses the PerL masking map to allow for precise control of pixel features for each object.

Table 2. Impact of integrating the PerL masking map (Object) into MagicDrive. We present the 3D object detection results based on BEVFormer [6], with outcomes showing superior performance emphasized in **bold**.

Method	FID↓	mAP↑	NDS↑	mAOE↓	mAVE↓	mATE↓
MagicDrive	**15.92**	15.21	28.79	0.81	0.57	0.95
MagicDrive + Mask	16.68	**15.54**	**29.77**	**0.73**	**0.56**	**0.89**

D.3. Effectiveness of View Cross-attention for Multi-View Consistency

View cross-attention ensures the seamless integration of visual data by maintaining continuity and consistency across the multiple camera feeds that are integral to current multi-functional perception systems in autonomous vehicles. Typically, autonomous vehicles feature a 360-degree horizontal surround view from a BEV perspective, resulting in overlapping fields of vision between adjacent cameras. Consequently, we facilitate direct interaction between the noise maps of each camera and those of the immediate left and right cameras. Given the noisy images from the current, left, and right cameras, designated as $\mathbf{Z}_b$, $\mathbf{Z}_l$, and $\mathbf{Z}_r$, respectively, the output of this multi-view generation is



w/ View Cross-attention

w/o View Cross-attention

Figure 3. Comparative visualization of outputs with (Top) and without (Bottom) view cross-attention. Red markers highlight discontinuities in the images generated without view cross-attention.

Algorithm 1 PerL-based Controlling Module (PerL-CM)

```
Input: road map features $\mathbf{H}_m \in \mathbb{R}^{1\times C}$, road masking map $\mathcal{M}_s \in \mathbb{R}^{HW\times 1}$, box features $\mathbf{H}_b \in \mathbb{R}^{M\times C}$, box masking map $\mathcal{M}_b \in \mathbb{R}^{HW\times M}$, scene text description features $\mathbf{H}_d \in \mathbb{R}^{1\times C}$, noisy multi-view street image feature $\mathbf{Z} \in \mathbb{R}^{HW\times C}$, and dimension $d$ (omit the detail of multi-view perspectives)
Output: Updated $\mathbf{Z}$
$\mathcal{A}_s \leftarrow softmax(\lambda_s \cdot \mathcal{M}_s + \mathbf{Z}\mathbf{H}_m^T/\sqrt{d})$
    // compute attention map for the road map in PerL-based cross-attention (scene)
$\mathbf{Z_s} \leftarrow \gamma_s \cdot \mathcal{A}_s\mathbf{H}_m + \mathbf{Z}$
$\mathcal{A}_b \leftarrow softmax(\lambda_b \cdot \mathcal{M}_b + \mathbf{Z_s}\mathbf{H}_b^T/\sqrt{d})$
    // compute attention map for the box in PerL-based cross-attention (object)
$\mathbf{Z_b} \leftarrow \gamma_b \cdot \mathcal{A}_b\mathbf{H}_b + \mathbf{Z_s}$
$\hat{\mathbf{Z}} \leftarrow \mathbf{Z}_b + \mathcal{C}(\mathbf{Z}_b, \mathbf{Z}_l, \mathbf{Z}_l) + \mathcal{C}(\mathbf{Z}_b, \mathbf{Z}_r, \mathbf{Z}_r)$
    // maintain visual consistency via View cross-attention
$\mathbf{Z}^* \leftarrow softmax(\hat{\mathbf{Z}}\mathbf{H}_d^T/\sqrt{d})\mathbf{H}_d + \hat{\mathbf{Z}}$
    // alter illumination and atmospheric effects by Text cross-attention
```

given by:

$$\hat{\mathbf{Z}} = \mathbf{Z}_b + \mathcal{C}(\mathbf{Z}_b, \mathbf{Z}_l, \mathbf{Z}_l) + \mathcal{C}(\mathbf{Z}_b, \mathbf{Z}_r, \mathbf{Z}_r), \tag{4}$$

where $\mathcal{C}(\cdot)$ represents the standard cross-attention operation, which accepts three input parameters: query, key, and value, respectively. This approach systematically integrates spatial information from various viewpoints, enabling the synthesis of images that exhibit visual consistency across distinct camera perspectives. Fig.3 offers a visual comparison of the model output with and without the application of view cross-attention. Upon integrating view cross-attention into PerLDiff, the procedure of the PerL-CM is detailed in Algo. 1.

Table 3. Ablation study comparing PerLDiff with a ControlNet-based model. We present 3D object detection results based on BEVFormer, BEVFusion. Outcomes demonstrating superior performance are highlighted in **bold**.

Method	Detector	FID↓	mAP↑	NDS↑	mAOE↓
ControlNet-based	BEVFormer	20.46	18.07	28.48	0.87
GLIGEN-based		**13.36**	**25.10**	**36.24**	**0.72**
ControlNet-based	BEVFusion	20.46	10.45	15.29	0.89
GLIGEN-based		**13.36**	**15.24**	**24.05**	**0.78**

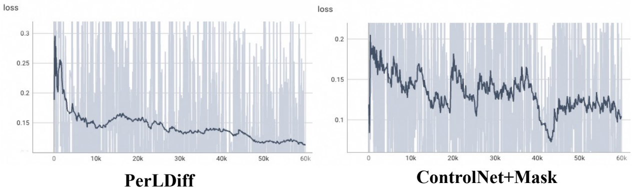


Figure 4. Training curves of PerLDiff and ControlNet-based, illustrating that Ours converges more rapidly during training.

D.4. Effectiveness of PerLDiff Based on ControlNet

In Tab. 3, we present an ablation study that replaces the architecture of PerLDiff with a ControlNet-based [14] model trained only on view cross-attention in Stable Diffusion. As shown in Tab.3, the performance of the ControlNet-based model is inferior to that of PerLDiff. Furthermore, Fig. 4 illustrates that PerLDiff employs a network architecture similar to GLIGEN [5], allowing it to converge more quickly on smaller datasets, such as NuScenes, compared to the ControlNet-based.

D.5. Effectiveness of Classifier-Free Guidance Scale

In Tab. 4, we assess the effect of the CFG [2] scale on the sampling of data generation. The term "scale" refers to the CFG scale, which is adjusted to balance conditional and unconditional generation. The transition from Method (b) to (e) indicates an increase in the CFG scale from 5.0 to 12.5. The results show an average increase of 2.87 in FID, an average decrease of 0.87% in mAP, an average reduction of 1.03% in NDS, a 0.02% increase in mAOE and a 1.07% drop in Vehicle mIoU. This provides substantial evidence that an excessively large CFG scale can degrade the quality of generated images and adversely affect various performance metrics.

E. Visualization Results

To further demonstrate the controllable generation capabilities of PerLDiff, we present additional visual results. Fig. 5 offers extended examples illustrating the superiority of PerLDiff in scene controllability, while Fig. 7 highlights

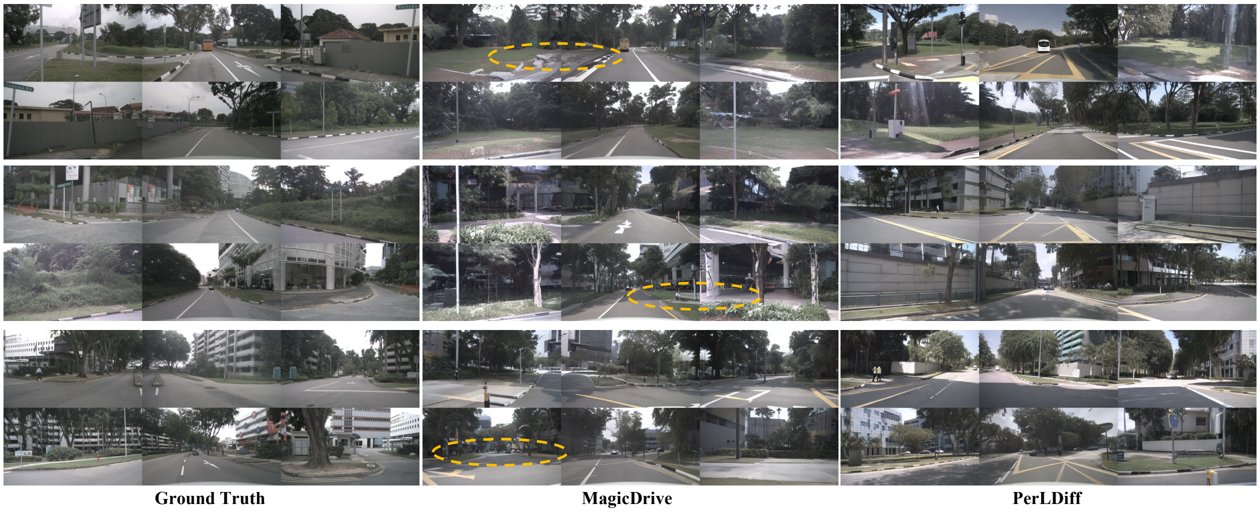


Figure 5. Qualitative comparison with MagicDrive. For **scene controllability**, PerLDiff demonstrates superior performance by results consistent with ground truth road information. Regions highlighted by yellow circles indicate areas where fail to align with ground truth.

Table 4. Comparison of different CFG [2] scale to each metric. We report the 3D object detection results based on BEVFormer [6] and BEV Segmentation results based on CVT [16].

Method	scale	FID↓	mAP↑	NDS↑	mAOE↓	Road mIoU↑	Vehicle mIoU↑
Oracle	–	–	27.06	41.89	0.54	70.35	33.36
(a)	2.5	**12.36**	23.89	36.03	**0.70**	60.05	26.95
(b)	5.0	13.36	**25.10**	**36.24**	0.72	61.26	**27.13**
(c)	7.5	15.52	24.62	35.60	0.74	**61.52**	26.63
(d)	10.0	16.32	24.20	35.05	0.73	61.43	26.00
(e)	12.5	16.86	23.86	34.98	0.74	61.25	25.55

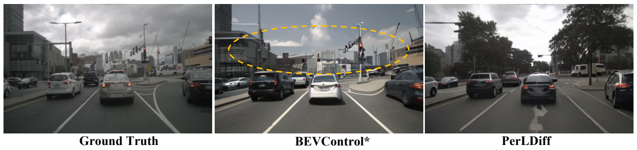


Figure 6. Qualitative results of the generated images reveal discrepancies in background details. As indicated by the yellow circle, PerLDiff produces background elements that do not align with real images due to the incorporation of the PerL masking map.

its effectiveness in controlling object orientation. Fig. 8 reveals that BEVControl* produces chaotic and indistinct attention maps leading to suboptimal controllability, PerLDiff optimizes the response areas of the attention map, resulting in accurate object-level control.

Table 5. Comparison with Panacea on synthetic 256×704 validation data. The quantitative evaluation using 3D object detection metrics from StreamPETR [11].

Method	FID↓	mAP↑	NDS↑	mAOE↓	mATE↓	mAVE↓
Oracle	–	47.05	56.24	0.37	0.61	0.27
Penacea	16.96	22.50	36.10	0.73	–	0.47
PerLDiff	**15.67**	**35.09**	**44.19**	**0.64**	**0.75**	**0.45**

Additionally, it is worth noting that, based on our experimental results as shown in Tab. 5, the key for temporal-based detection models lies in accurately positioning and categorizing objects in each frame; detailed information about objects, such as color and brand, is not crucial. As illustrated in Fig. 9, when provided with continuous frame inputs, the generated images by PerLDiff ensure that the positions and categories of objects, along with the road map, are consistently aligned with the specified conditions between adjacent frames.

Moreover, Fig. 10 displays scene alterations by PerLDiff to mimic different weather conditions or times of day, showcasing its versatility in changing scene descriptions. Furthermore, as illustrated in Fig. 6, PerLDiff generates background details that do not fully align with those of real images. This discrepancy arises because PerLDiff incorporates prior constraints to ensure accuracy in object detection, which can, in turn, negatively impact the fidelity of the background details.

Finally, Fig. 11 presents samples from KITTI *validation* set, illustrating the application's performance in real-world conditions.

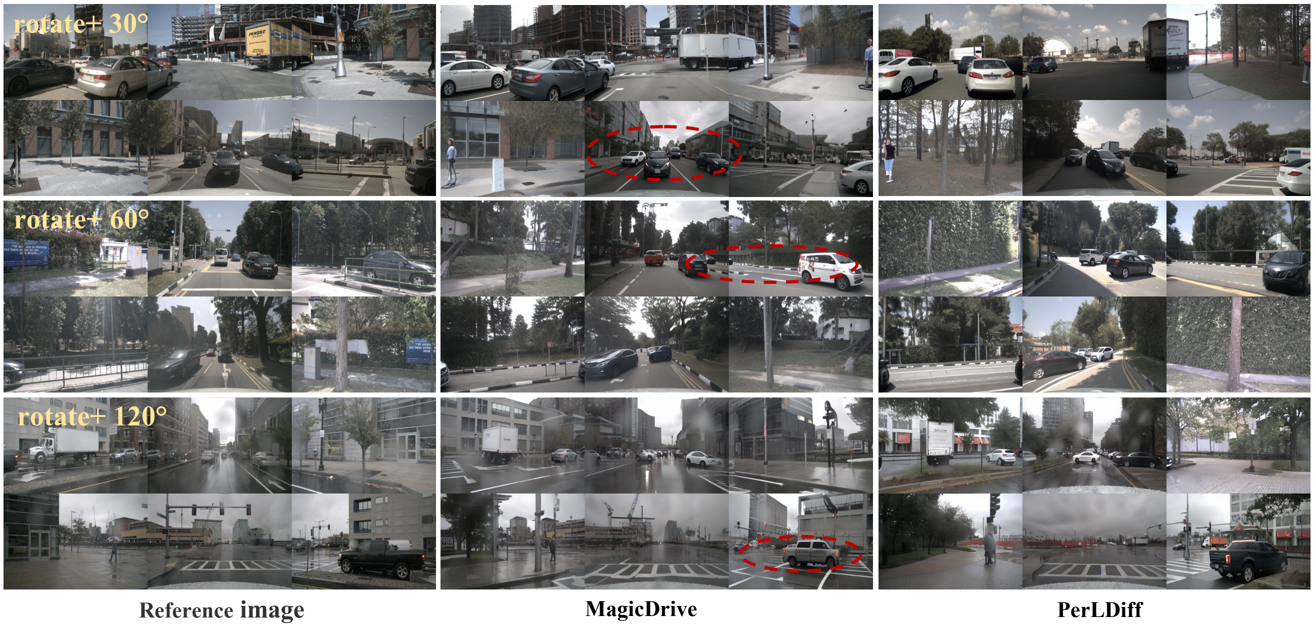


Figure 7. Qualitative comparison with MagicDrive. For **object controllability**, PerLDiff exhibits superior performance by generating objects at arbitrary angles. Regions highlighted by red circles denote scenarios where the images fail to achieve correct orientation.

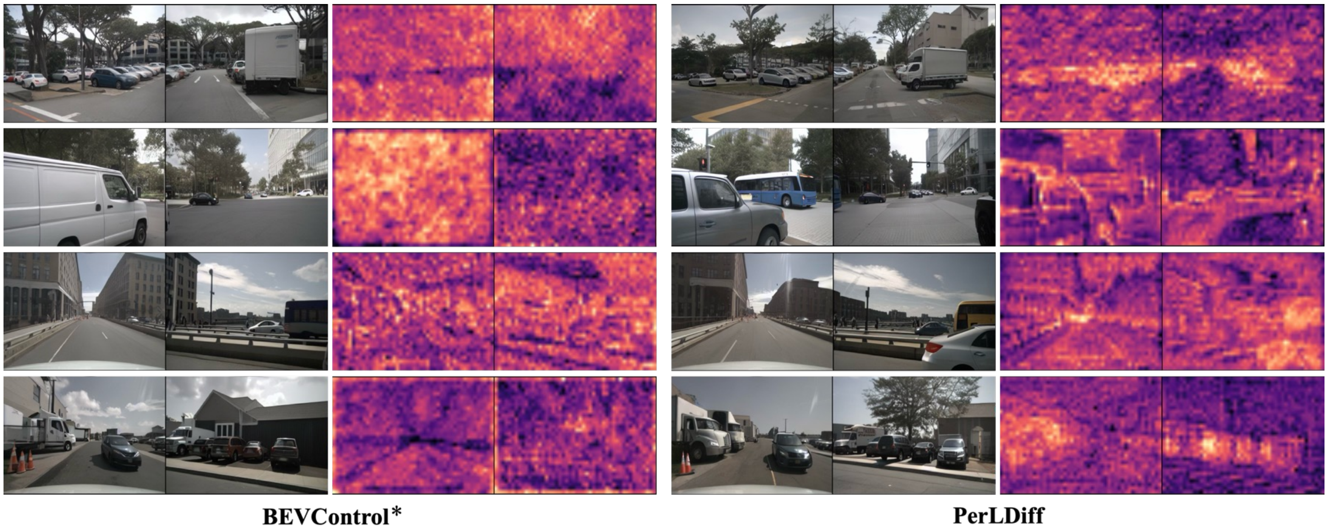


Figure 8. Visualization of cross-attention map results. From left to right, we present the generated images and corresponding cross-attention maps from our baseline BEVControl* and our PerLDiff. BEVControl* produces disorganized and vague attention maps, which result in inferior image quality. Conversely, PerLDiff method fine-tunes the response within the attention maps, resulting in more accurate control information at the object level and improved image quality.

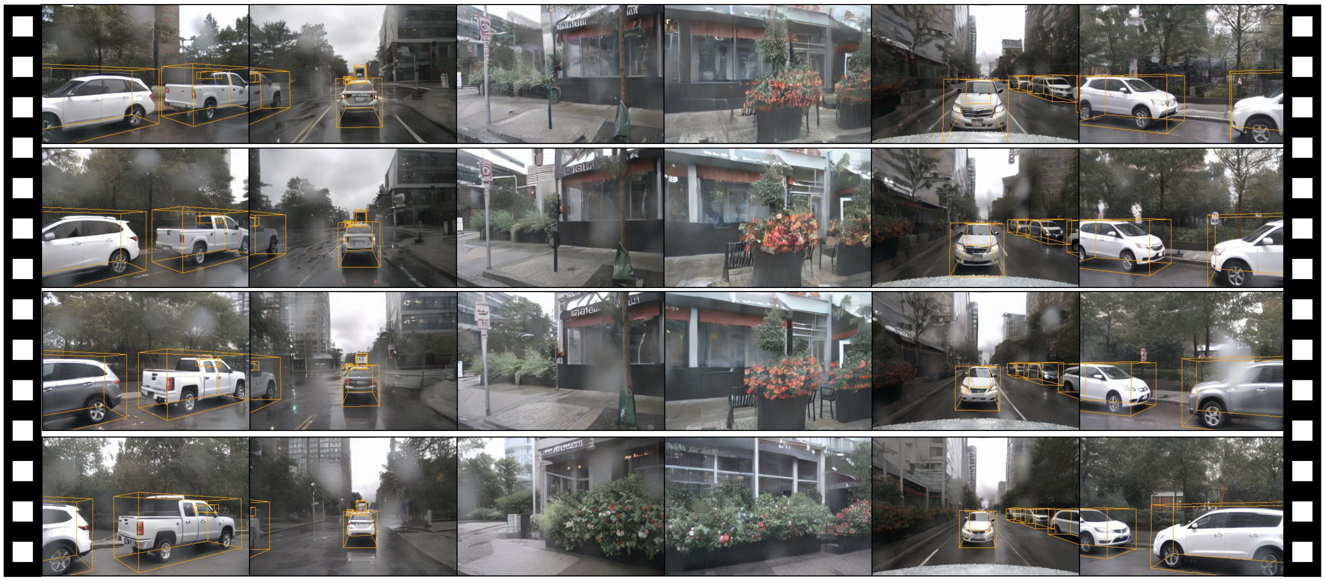


Figure 9. Qualitative visualizations from the NuScenes. PerLDiff demonstrate consistent alignment of object positions and categories, along with the road map, when provided with continuous frame inputs, ensuring coherence between adjacent frames.

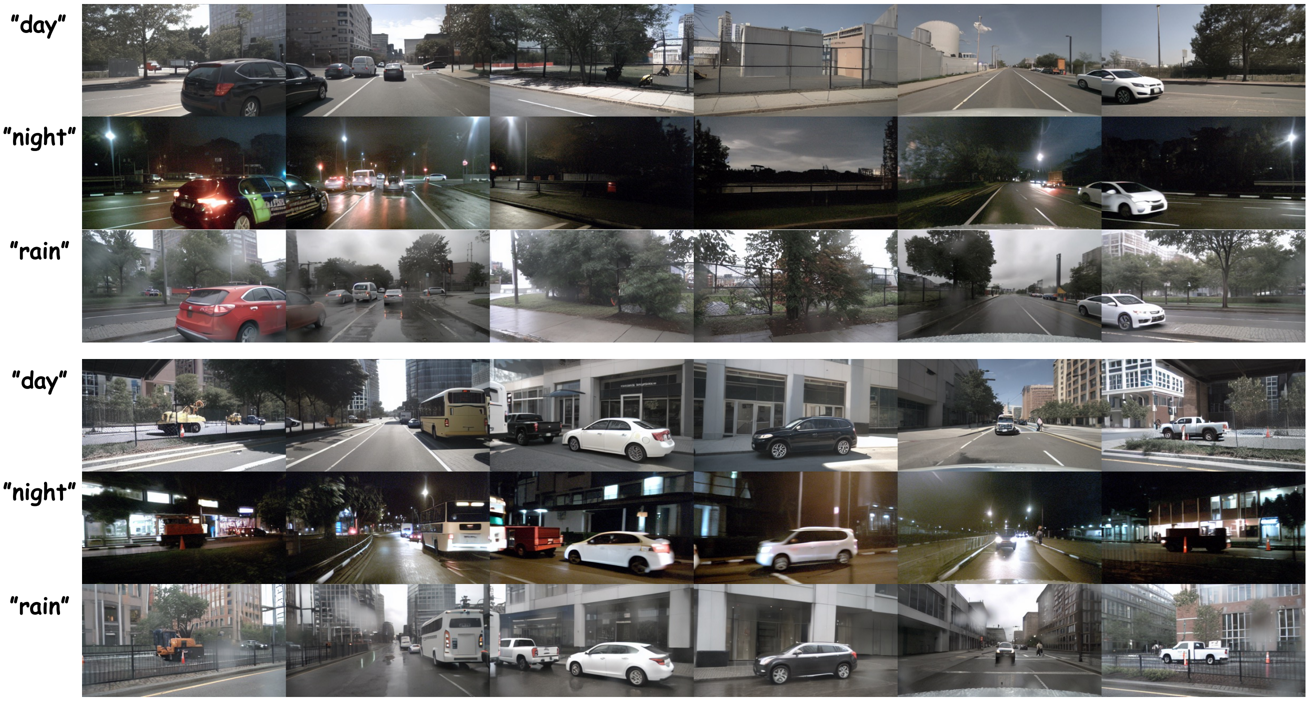


Figure 10. Qualitative visualization on NuScenes demonstrating the effects of Text Cross-attention. From top to bottom: *day*, *night*, and *rain* scenarios synthesized by PerLDiff, highlighting its adaptability to different lighting and weather conditions.

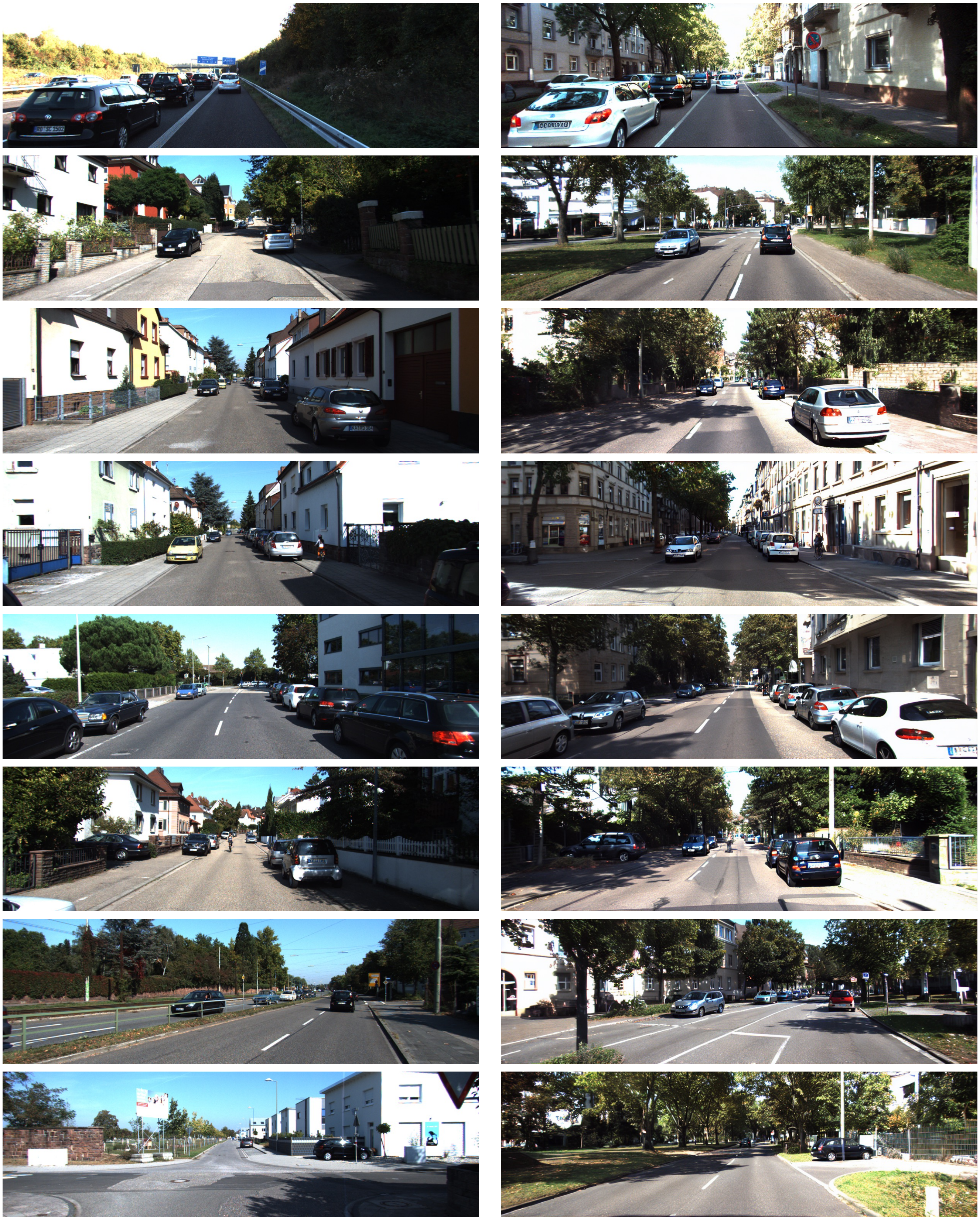


Figure 11. Visualization of street view images generated by our PerLDiff on KITTI *validation* dataset. We show the ground truth (left) and our PerLDiff (right).

References

[1] 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. *arXiv preprint arXiv:2310.02601*, 2023. 2

[2] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 3, 4

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[4] 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. 1

[5] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 3

[6] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1, 2, 4

[7] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 1, 2

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[11] 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. 1, 4

[12] 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. *arXiv preprint arXiv:2311.17918*, 2023. 1

[13] 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. *arXiv preprint arXiv:2311.16813*, 2023. 1

[14] 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*, pages 3836–3847, 2023. 3

[15] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3298, 2021. 1

[16] 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*, pages 13760–13769, 2022. 1, 4