NoiseController: Towards Consistent Multi-view Video Generation via Noise Decomposition and Collaboration
–Supplementary Material–

Haotian Dong[1,†], Xin Wang[2,†], Di Lin[1,†], Yipeng Wu[1], Qin Chen[1], Ruonan Liu[3,🖂]
Kairui Yang[1], Ping Li[2], and Qing Guo[4]
[1] Tianjin University [2] The Hong Kong Polytechnic University
[3] Shanghai Jiao Tong University [4] Nankai University

Contents

List of Figures

[†] Haotian Dong, Xin Wang, and Di Lin are co-first authors.
[🖂] Ruonan Liu is the corresponding author.

List of Tables

1. Theoretical Foundations for NoiseController

Our multi-level noise decomposition does not violate the principles of Diffusion Models (DMs). The original DMs take whole-image Gaussian noise as inputs, which inherently limits their controllability in local regions. Many studies like [4, 7, 10] facilitate precise control in local regions by modifying the initial noise through masking, disrupting the global Gaussian distribution of the initial noises. Specifically, Mao et al. [4] mention that diffusion models can tolerate non-Gaussian noise. FreeInit [10] decomposes noise into different-frequency components and modifies the high-frequency noise. This approach also results in a change of the initial noise distribution. To adapt the diffusion model to non-Gaussian noise, existing works [5] [1] fine-tune the decoder of DMs, achieving better generation performance and enabling the precise control in local regions.

Notably, from a technical perspective, previous works [11, 12] typically predict the mean and variance of the noise, and then sample the noise for progressive denoising. However, from the perspective of neural network fitting, it is impossible to ensure that the estimated noise perfectly aligns with the added Gaussian noise when training the noise estimation model. Current works [1, 5] directly predict the t^{th} step noise by training the diffusion model, allowing for direct supervision of the predicted noise against the ground-truth added noise. This strategy facilitates the model's adaptation to non-Gaussian noise through fine-tuning.

Guided by this principle, ***NoiseController*** performs multi-level noise decomposition and multi-frame noise collaboration, to compute the consistent ground-truth noises using scene-level 6-view masks, enabling separate controlled generation of local background and foreground regions. Finally, we enhance the consistency and quality of multi-view video generation by fine-tuning the diffusion model to adapt to this consistent initial noise.

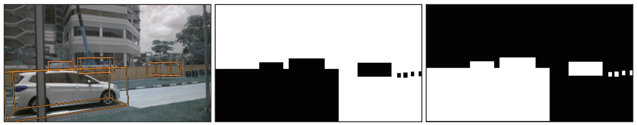
2. Calculation of Scene-level Mask

We illustrate the calculation of the scene-level foreground and background masks in Figure 1. The scene-level background mask $\mathbf{M}^{\mathrm{B}}$ is derived from the bounding box, where the pixel within the box region is 0, otherwise is 1. The foreground mask $\mathbf{M}^{\mathrm{F}}$ is the inverse of $\mathbf{M}^{\mathrm{B}}$. We define the calculation of scene-level foreground mask $\mathbf{M}^{\mathrm{F}}$ as:

$$\mathbf{M}^{\mathrm{F}} = 1 - \mathbf{M}^{\mathrm{B}} \tag{1}$$

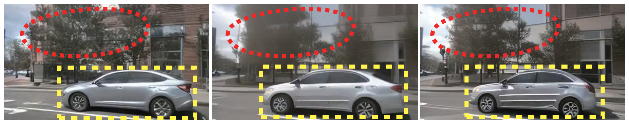
3. Additional Ablation Studies

We provide additional ablation studies for NoiseController to demonstrate the effectiveness of its core modules.



(a) Image **(b) Background Mask $\mathbf{M}^{\mathrm{B}}$** **(c) Foreground Mask $\mathbf{M}^{\mathrm{F}}$**

Figure 1. We compute the scene-level background and foreground masks based on the bounding box.



(a) NoiseController **(b) Only U-Net$_{\mathbf{F}}$** **(c) Only U-Net$_{\mathbf{B}}$**

Figure 2. The visualization comparison of generated video frame via joint denoising and independent denoising. (a) NoiseController achieves high-quality generation performance on both background and foreground regions, surpassing independent denoising, e.g., (b) only using U-Net$_{\mathbf{F}}$ and (c) U-Net$_{\mathbf{B}}$.

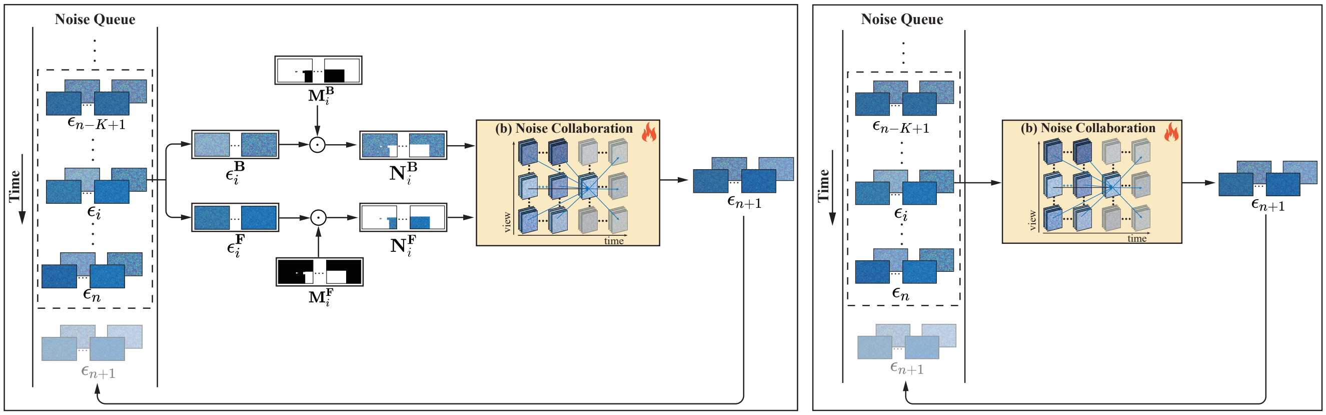
3.1. Independent Results of U-Net$_{\mathbf{B}}$ and U-Net$_{\mathbf{F}}$

We employ our NoiseController that incorporates joint denoising to achieve high-quality generation performance in both background and foreground regions (see Figure 2(a)). Independent denoising using U-Net$_{\mathbf{F}}$ (see red ellipse in Figure 2(b)) or U-Net$_{\mathbf{B}}$ (see yellow rectangle in Figure 2(c)) results in deficient performance in background or foreground regions, respectively. We observe that independent denoising using U-Net$_{\mathbf{F}}$ achieves superior generation performance for foreground objects but struggles to generate high-quality background scenes, compared to the result of U-Net$_{\mathbf{B}}$. This is because that joint denoising alternates between foreground and background, preventing low-quality components from affecting each other.

3.2. Combinations of Scene-Level Noise Mask and Multi-Frame Noise Collaboration

We conduct experiments on the effect of multi-level noise decomposition and multi-frame noise collaboration orders. The performances are reported in Table 1 in terms of FVD and FID. As illustrated in Figure 3 (a), we leverage scene-level 6-view noises $\epsilon_i^{\mathbb{D}}$ and corresponding scene-level 6-view masks $\mathbf{M_i^{\mathbb{D}}}$ to compute the masked background noise $\mathbf{N}_i^{\mathbf{B}}$ and masked foreground noise $\mathbf{N}_i^{\mathbf{F}}$ for the i^{th} frame. Then, we leverage the inter-view spatiotemporal collaboration matrix $\mathbb{S}$ and intra-view impact collaboration matrix $\mathbb{I}$ to compute the following scene-level 6-view noises, ensuring spatiotemporal consistency of the initial 6-view N-frame noises. Compared to the NoiseController that masks the computed 6-view noises $\epsilon_n^{\mathbb{D}}$ with the scene-level 6-view masks $\mathbf{M_n^{\mathbb{D}}}$ after multi-frame noise collaboration (as illustrated in Figure 3 (b), it degrades the performances into FVD 170.3 and FID 26.01 (see the first row of Table 1).

Decomposing the initial noises into background and



(a) Mask-Collaboration

(b) Collaboration-Mask

Figure 3. (a) Detailed architecture of Mask-Collaboration. The scene-level background noises $\epsilon_i^{\mathbf{B}}$ and foreground noises $\epsilon_i^{\mathbf{F}}$ are masked with scene-level masks $\mathbf{M}_i^{\mathbf{B}}$ and $\mathbf{M}_i^{\mathbf{F}}$ for spatial decomposition. Then, the masked noises $\mathbf{N}_i^{\mathbf{B}}$ and $\mathbf{N}_i^{\mathbf{F}}$ are fed into multi-frame noise collaboration to compute the 6-view noises ϵ_{n+1} for the $(n+1)^{th}$ frame. (b) Detailed architecture of Collaboration-Mask. We respect the preceding K-frame information, leveraging multi-frame noise collaboration to compute the 6-view noises ϵ_{n+1} for the $(n+1)^{th}$ frame.

Table 1. We examine the effectiveness of multi-level noise decomposition and multi-frame noise collaboration. Results are reported regarding FVD and FID on the validation set of nuScenes.

Method	FVD ↓	FID ↓
Mask-Collaboration	170.3	26.01
Collaboration-Mask	**122.9**	**14.65**

foreground noises fails to capture the internal collaborations between scene-level background and foreground noises, resulting in poor consistency in noise level. The inconsistent 6-view N-frame noises further exacerbate the inconsistencies during generation, which leads to unsatisfactory consistency in the generated video frames. The coupled background and foreground information naturally maintain the tight connection, which provides the scene-level information to enhance spatiotemporal consistency and realism in generated videos. It should be noted that NoiseController decomposes the 6-view N-frame initial noises into scene-level background and foreground noises without spatial decomposition before noise collaboration, ensuring the multi-frame noise collaboration to model the internal collaborations between these noises. After the estimated 6-view noises $\tilde{\epsilon}^{\mathbb{D}}_{n+1}$ via multi-frame noise collaboration, we respect the scene-level masks to decompose the background and foreground spatially, which allows two parallel denoising U-Nets to focus on the generation of different spatial distributions.

3.3. The Setting of the Initial Noise

We conduct experiments on the setting of initial noise. The performances are reported in Table 2 in terms of FVD and FID. We provide unsuited noise prior for denoising network, when we leverage the same noises across 6 views within the same frame, leading to worse performance (see the first row of Table 2). We then utilize 6-view N-frame different noises for multi-view video generation. In this scenario, our NoiseController degrades into the baseline model, which fails to adequately constrain the initial noises, resulting in inconsistent starting points for denoising and the inferior generation performance (see the second row of Table 2). Our NoiseController captures mutual cross-view effects and historical cross-frame impacts through multi-level noise decomposition and multi-frame noise collaboration, enhancing the consistency of generated multi-view videos (see the last row of Table 2).

3.4. Weights of Shared and Residual Components of Scene-Level Noises

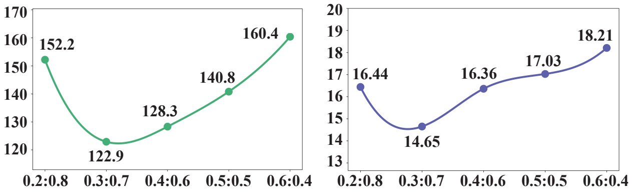
In this paper, each residual component of scene-level noises follows a specific Gaussian distribution, i.e., the residual components of scene-level noise $\epsilon_{m,n}^{\mathbf{B}_\mathrm{R}}$ follows the Gaussian distribution of $\mathcal{N}(\mathbf{0}, \frac{1}{\eta^2+1}\mathbf{I})$ and $\epsilon_{m,n}^{\mathbf{F}_\mathrm{R}}$ follows $\mathcal{N}(\mathbf{0}, \frac{1}{\lambda^2+1}\mathbf{I})$, where η and λ are two hyperparameters. We conduct experiments on the effectiveness of shared component weights $\mathbf{W}^{\mathbb{D}_\mathrm{S}}$ and residual components weight $\mathbf{W}^{\mathbb{D}_\mathrm{R}}$

Table 2. We conduct experiments on different initial noise settings. We randomly sample the noise for 6 views for a specific video frame, where 6-view noises are same (see **SameNoise**) or different (see **DifferentNoise**). Results are reported regarding FVD and FID on the validation set of nuScenes.

Noise	FVD ↓	FID ↓
SameNoise	234.8	23.54
DifferentNoise	177.3	20.92
NoiseController	**122.9**	**14.65**

Table 3. Application of using the generated data to augment the downstream tasks. The performances are evaluated on the nuScenes validation set.

Data	**Detection**						**Segmentation**	
	BEVDet [3]		BEVDepth [6]		StreamPETR [9]		FIERY [2]	Lift [8]
	NDS ↑	mAP ↑	NDS ↑	mAP ↑	NDS ↑	mAP ↑	IoU ↑	IoU ↑
w/o aug.	35.00	28.30	43.55	33.04	53.70	43.20	35.80	33.03
w/ NoiseController	**36.02**	**28.92**	**44.09**	**33.21**	**53.92**	**43.32**	**37.30**	**33.89**



(a) FVD

(b) FID

Figure 4. Comparisons on the impact of the weights of the shared and residual components of scene-level noises. We report the performances in terms of FVD and FID on the validation set of nuScenes.

of scene-level background and foreground noises. The performances are reported in Figure 4, where abscissa is $\mathbf{W}^{\mathbb{D}_\mathrm{S}} : \mathbf{W}^{\mathbb{D}_\mathrm{R}}$.

Too small shared weights, e.g., $\mathbf{W}^{\mathbb{D}_\mathrm{S}} : \mathbf{W}^{\mathbb{D}_\mathrm{R}} = 0.2 : 0.8$, NoiseController achieves the unsatisfactory performances with FVD 152.2 and FID 16.44 because a small proportion of shared components decreases the multi-frame noise collaboration, resulting in unsatisfactory consistency cross frames.

With larger shared weights, e.g., $\mathbf{W}^{\mathbb{D}_\mathrm{S}} : \mathbf{W}^{\mathbb{D}_\mathrm{R}} = 0.3 : 0.7$, NoiseController improves the performances to FVD 122.9 and FID 14.65. The larger shared component weights help to collaborate cross-frame noises, which ensures the computation of shared components of scene-level noises fully utilizes the preceding K-frame information.

However, too large shared weights, e.g., $\mathbf{W}^{\mathbb{D}_\mathrm{S}} : \mathbf{W}^{\mathbb{D}_\mathrm{R}} = 0.6 : 0.4$, performances degrade significantly with FVD 160.4 and FID 18.21. This is because too large shared component weights result in content homogenization, resulting in a lack of diversity in generated video frames. Therefore, we choose $\mathbf{W}^{\mathbb{D}_\mathrm{S}} : \mathbf{W}^{\mathbb{D}_\mathrm{R}} = 0.3 : 0.7$ as the default.

4. Migration of Consistent Noises

Our NoiseController provides spatiotemporally consistent initial noises through multi-level noise decomposition and multi-frame noise collaboration. We transfer the foreground and background noises generated by MagicDrive into NoiseController, respectively, to examine the effectiveness of our designed consistent noise. As shown in Figure 5, replacing our consistent noises with foreground and background noises generated by MagicDrive results in the loss of consistency in the generated video frames. As shown in Figure 6, replacing MagicDrive's foreground and background noises with our consistent noises leads to the addition of consistency in the generated video frames.

5. Extra Downstream Applications

We interpolate the frames in each scene of the original nuScenes dataset (2 Hz, 28,130 samples for training, and 6,019 samples for validation) to achieve a sampling frequency of 12 Hz. In each scene, we segment the data into video clips using a sliding window whose length is 16 frames, moving forward with a stride of one frame. It allows us to generate 154,780 and 33,114 video clips for training and validation sets, respectively.

We randomly perform video generation on 2% of the videos based on the training set. Then, we select the images according to the sample tokens from the original nuScenes training set to achieve a large-scale dataset consisting of images generated by NoiseController. Given this large-scale dataset, we randomly select 5,600 video frames comprising 6-view images to serve as additional diverse training data for downstream perception tasks.

We choose detection and segmentation as the downstream perception tasks. Specifically, we choose BEVDet [3], BEVDepth [6], and StreamPETR [9] as the baseline downstream detection models, while FIERY [2] and Lift [8] are served as the baseline segmentation models. We retrain these methods on the mixed training set which consists of the original training set and additional training data generated by NoiseController, to evaluate the effectiveness of generating data for augmenting the perception tasks. The experimental setups including the training epochs and learning rate of each method are consistent with those in the original paper.

In Table 3, we report the evaluation performance without augmentation (see the first row of Table 3), as well as with augmentation using data generated by NoiseController (see the second row) for downstream detection and segmentation tasks. As a data augmentation strategy, these five methods have achieved a significant improvement by retraining on the mixed training set. This is because NoiseController fully utilizes multi-level noise decomposition and multi-frame noise collaboration to enhance the consistency

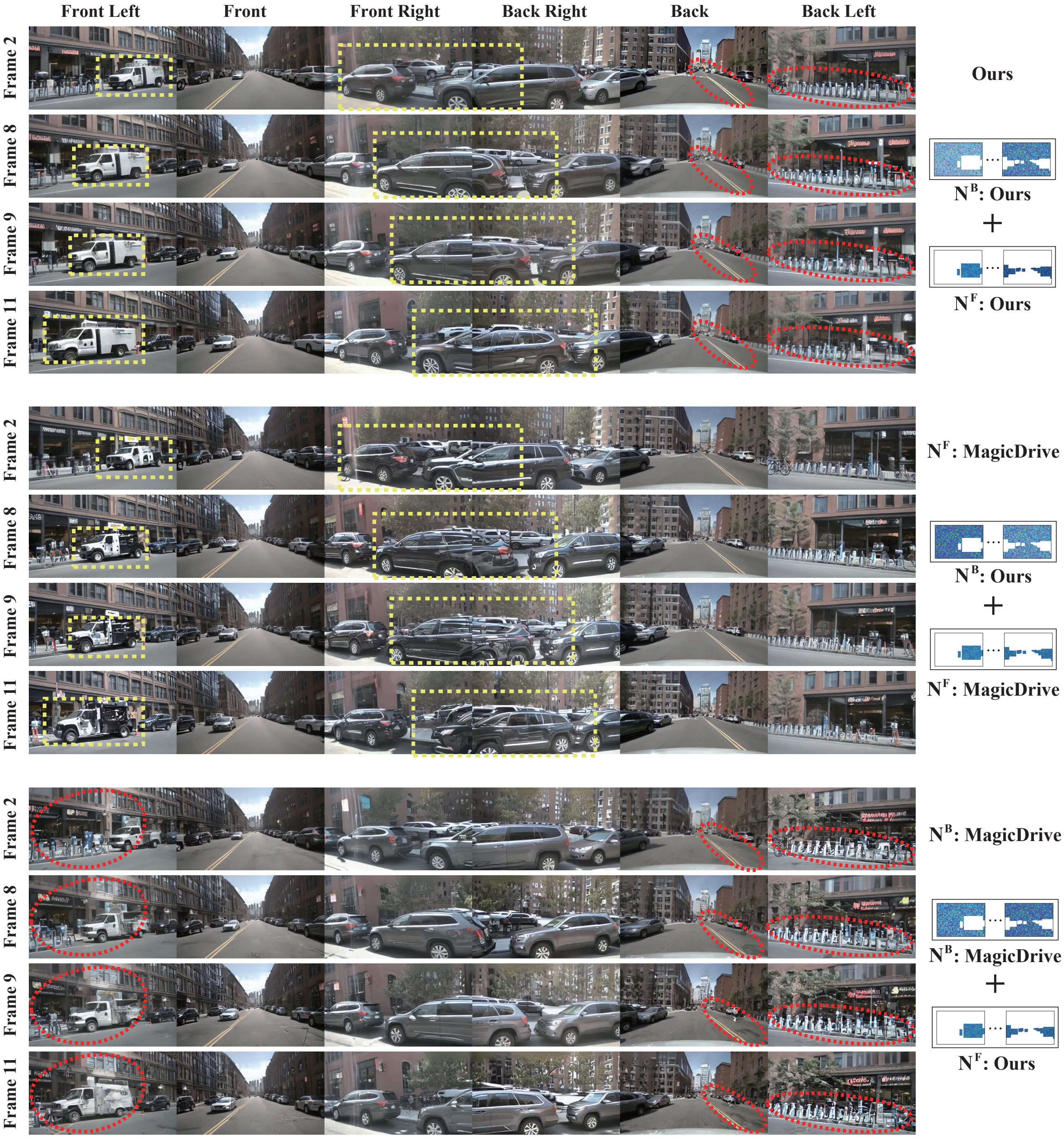


Figure 5. The visual comparisons on replacing our consistent foreground and background noises with those generated by MagicDrive. $\mathbf{N^B}$ and $\mathbf{N^F}$ represent the masked background and foreground noises, respectively. In the second case, we replace our foreground noise with MagicDrive's foreground noise, resulting in inferior performance on generated foreground objects (see yellow rectangles). In the third case, we use our foreground noise and MagicDrive's background noise to achieve the overall noise for video generation, leading to inferior performance on background scenes (see red ellipses).

in noise level, leading to spatiotemporal consistency in generated videos. The additional diverse training data generated by our NoiseController helps downstream detection and segmentation models to optimize network parameters during the training process and enhance the robustness of models.

6. Additional Visual Comparisons

In Figures 7, 8, 9, and 10, we provide more visualization comparisons on the validation set of nuScenes. Our inter-view spatiotemporal collaboration matrix $\mathbb{S}$ models the noise collaborations between different views, helping to enhance the cross-view consistency in the generated videos. As shown in Figure 7, NoiseController maintains better cross-view consistency for foreground objects compared to the MagicDrive. In multi-frame noise collaboration, we respect the preceding 6-view K-frame noises to compute the shared components of scene-level noises for the following frames, helping to enhance the cross-frame consistency in the generated video frames. Please note that NoiseController can maintain considerable cross-frame consistency for foreground objects (see Figure 8) as well as background scenes (see Figure 9). We provide the generated videos for better comparisons on cross-view and cross-frame consistency. In Figure 10, we show the comparisons of the details of background scenes and foreground objects. The spatial decomposition of scene-level background and foreground noises ensures that two parallel denoising U-Nets of NoiseController focus on the different spatial distributions, enhancing the generation performance on background and foreground details. Additionally, we include some video results in "video.zip".

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. In *ICLR*, 2024. 1

[2] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *ICCV*, pages 15273–15282, 2021. 3

[3] Junjie Huang, Guan Huang, Zheng Zhu, Ye Yun, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 3

[4] Xueting Wang Jiafeng Mao and Kiyoharu Aizawa. The lottery ticket hypothesis in denoising: Towards semantic-driven initialization. In *ECCV*, 2024. 1

[5] Pengxiang Li, Kai Chen, Zhili Liu, Ruiyuan Gao, Lanqing Hong, Guo Zhou, Hua Yao, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Trackdiffusion: Tracklet-conditioned video generation via diffusion models. In *WACV*, 2025. 1

[6] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, pages 1477–1485, 2023. 3

[7] Jiafeng Mao, Xueting Wang, and Kiyoharu Aizawa. Guided image synthesis via initial image editing in diffusion model. In *ACM MM*, pages 5321–5329, 2023. 1

[8] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pages 194–210, 2020. 3

[9] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *ICCV*, 2023. 3

[10] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. In *ECCV*, pages 378–394, 2024. 1

[11] Ziyuan Zhong, Davis Rempe, Yuxiao Chen, Boris Ivanovic, Yulong Cao, Danfei Xu, Marco Pavone, and Baishakhi Ray. Language-guided traffic simulation via scene-level diffusion. In *RA-L*, pages 144–177, 2023. 1

[12] Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Veer, Tong Che, Baishakhi Ray, and Marco Pavone. Guided conditional diffusion for controllable traffic simulation. In *ICRA*, pages 3560–3566, 2023. 1

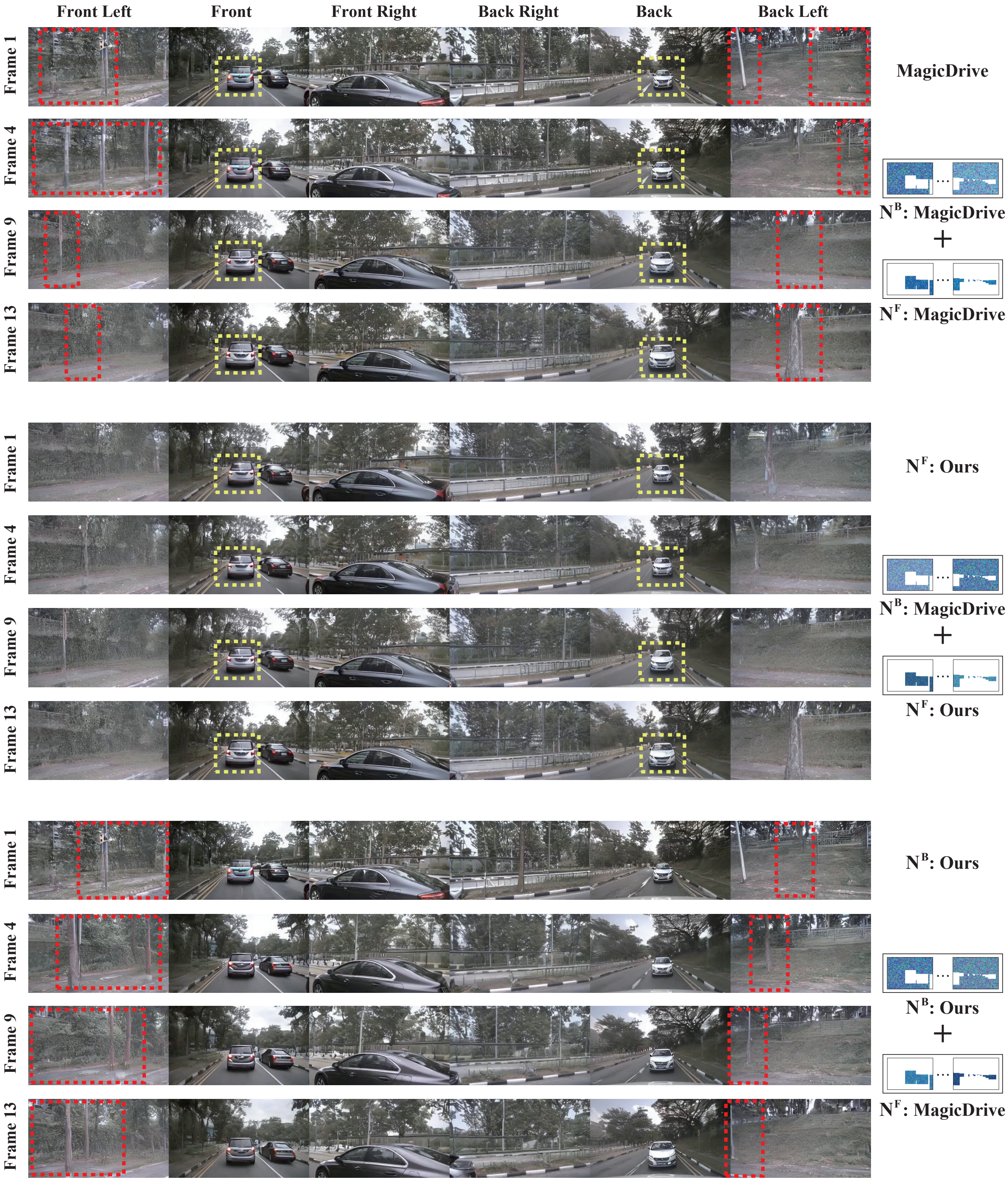


Figure 6. The visual comparisons on replacing Magicdrive's foreground and background noises with our consistent noise respectively. $\mathbf{N}^{\mathbf{B}}$ and $\mathbf{N}^{\mathbf{F}}$ represent the masked background noise and masked foreground noise, respectively. Integrating our consistent scene-level background and foreground noises into MagicDrive can significantly improve its performance.

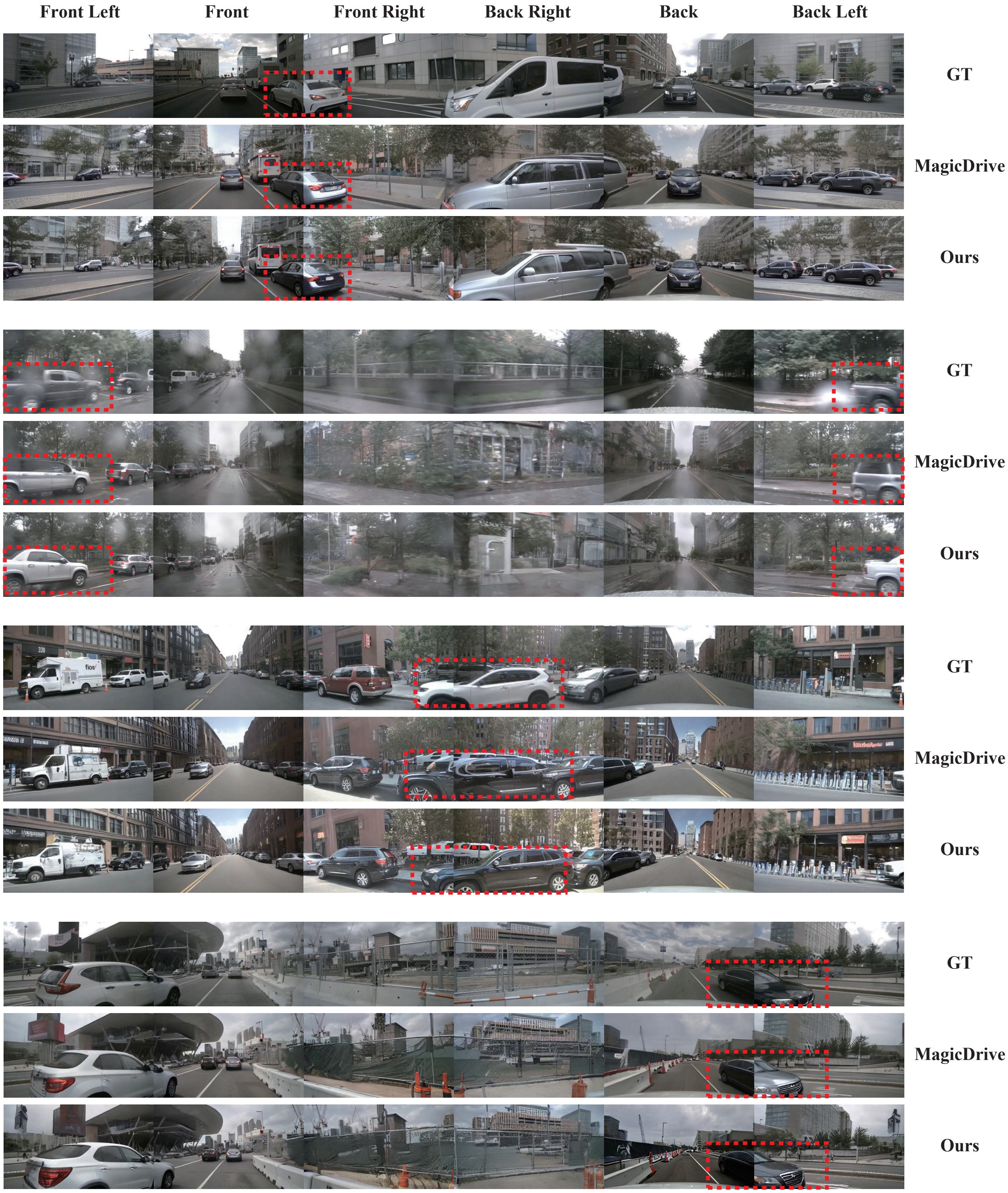


Figure 7. The visual comparisons on cross-view consistency between MagicDrive and NoiseController. We highlight the compared regions (see red rectangles) for better comparisons. Our NoiseController achieves better cross-view consistency in generated video frames.

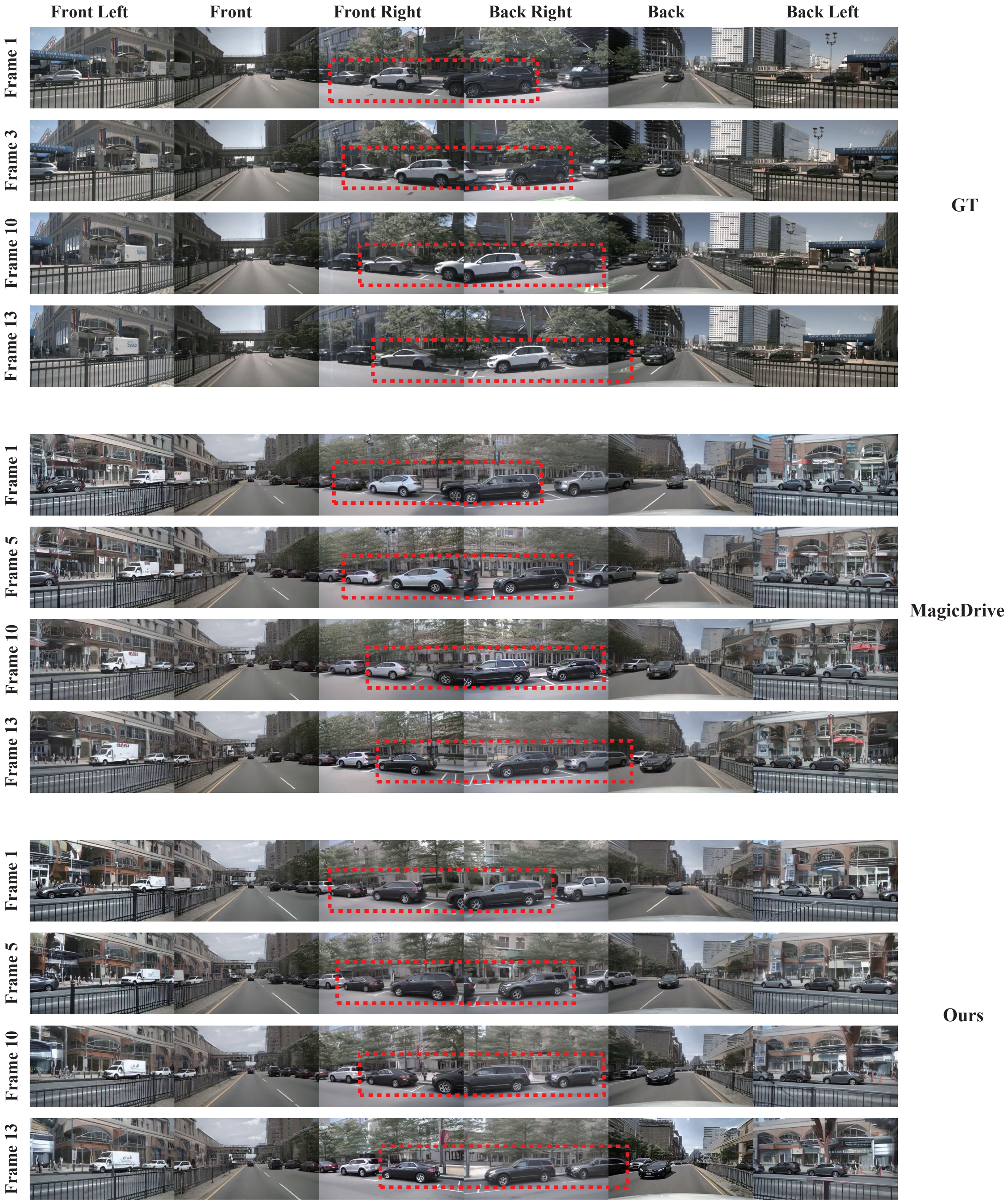


Figure 8. The visual comparisons on cross-frame consistency between MagicDrive and NoiseController. We visualize the 1^{st}, 5^{th}, 10^{th}, and 13^{th} video frames with highlighted regions (see red rectangles) for better comparison. Equipped with multi-level noise decomposition and multi-frame noise collaboration, our NoiseController achieves better cross-frame consistency for foreground objects in generated videos.

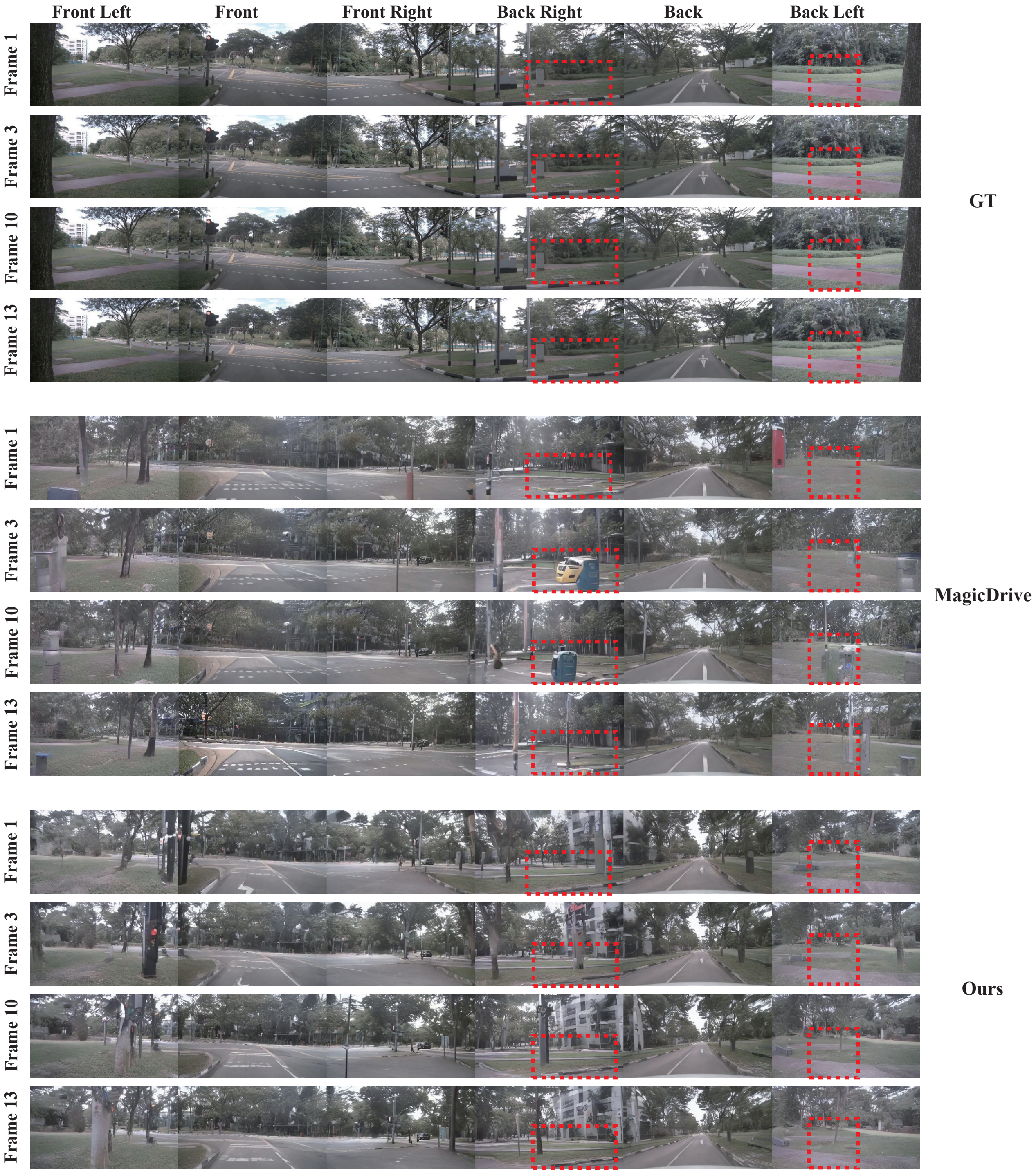


Figure 9. The visual comparisons on cross-frame consistency between MagicDrive and NoiseController. We visualize the 1^{st}, 3^{rd}, 10^{th}, and 13^{th} video frames with highlighted regions (see red rectangles) for better comparison. Our NoiseController achieves better cross-frame consistency for background scenes in generated videos due to the usage of multi-level noise decomposition and multi-frame noise collaboration.

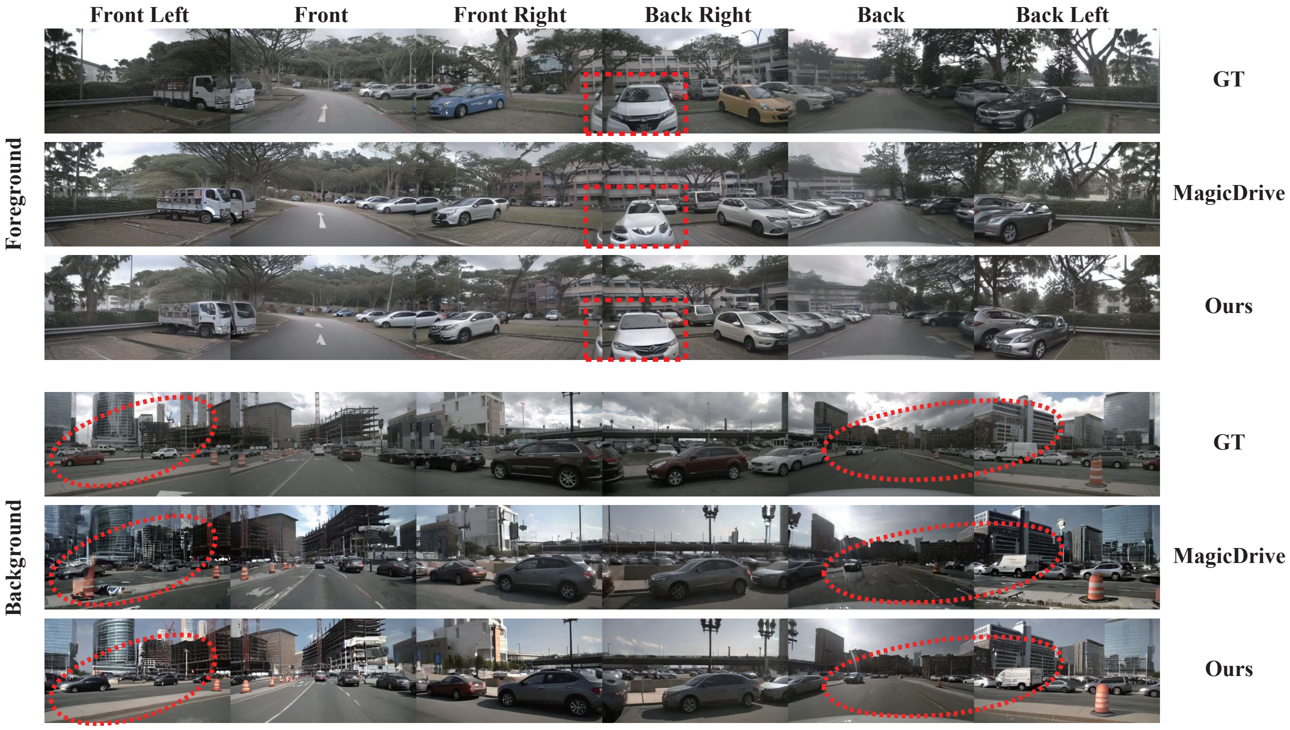


Figure 10. The visual comparisons of foreground and background details between MagicDrive and NoiseController. Decomposing the initial noises into scene-level background noises and foreground noises, our NoiseController can focus on different spatial distributions during joint denoising. It allows us to generate better background scenes and foreground objects along with detailed structures.