WAVE: Warp-Based View Guidance for Consistent Novel View Synthesis Using a Single Image (Supplementary Material)

Contents

Page

A. Code and Website 1

B. More Details of Methods 1
B.1. Validating the assumption of warp-guided adaptive Attention . . . 1

C. Metrics 2
C.1. Video consistency metrics . . . 2
C.2. Camera parameter accuracy . . . 3

D. Experiment Details 4
D.1. Implementation details . . . 4
D.2. Consistency and camera accuracy experiment . . . 4
D.3. Target view image reconstruction experiment . . . 4
D.4. RE10K sequence evaluation . . . 5
D.5. Downstream task . . . 5
D.6. Quantitative evaluation of adaptive warp-range selection effect . . . 6

E. More Qualitative Results 6
E.1. Camera parameter visualization . . . 6
E.2. Additional samples . . . 7

F. Failure Cases 12

A. Code and Website

The code for WAVE will be released soon. A project website has been created to introduce WAVE in a simple and accessible manner, as well as to showcase a variety of qualitative results and videos. It can be accessed via the following link: *project page:* https://jwoo-park0.github.io/wave.github.io/

B. More Details of Methods

B.1. Validating the assumption of warp-guided adaptive Attention

We propose **warp-guided adaptive attention** to manipulate attention, utilizing warped region masks. This approach is based on the hypothesis that the attention mechanism in the U-Net decoder layers preserves spatial position correspondence. Leveraging this assumption, we design warped region masks that align with the attention maps, enabling direct attention manipulation. This method efficiently incorporates the novel viewpoint information into the attention mechanism, contributing to generating consistent images.

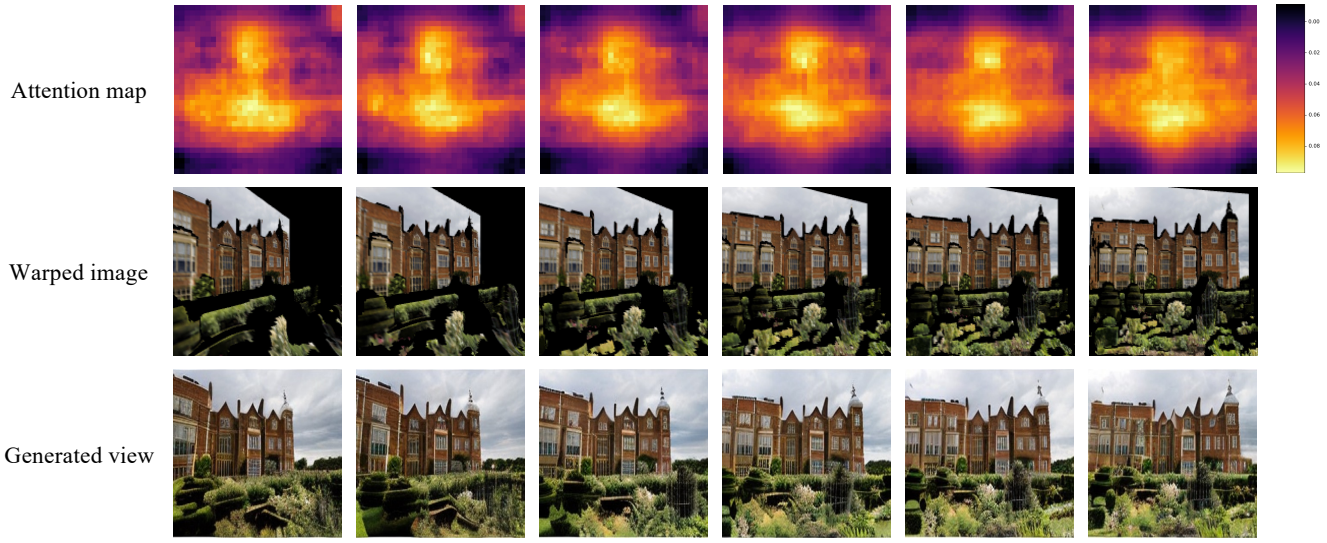

Figure 1. **Visualization of attention maps, warp images and generated views.** We present the attention maps of the model alongside the warp images to validate the assumptions of our proposed warp-guided adaptive attention. The results demonstrate that the attention map adapts in a similar manner to changes in the warp image.

To validate the hypothesis above, attention maps are extracted from the U-Net decoder layers. Attention maps are extracted at every DDIM [13] step, with those at different resolutions interpolated and averaged. Specifically, attention maps are examined to analyze how the attention distribution varies across different viewpoints. As demonstrated in Fig. 1, the attention maps dynamically adjust based on the warped images. This observation confirms our initial hypothesis that the decoder attention retains spatial position correspondence. Thus, our study substantiates the proposed approach.

C. Metrics

C.1. Video consistency metrics

We introduce **LPIPS-*first***, **CLIPSIM-*first***, **LPIPS-*next***, and **CLIPSIM-*next***, the metrics proposed in previous works [3, 23] to evaluate video consistency. Originally designed to assess frame-to-frame consistency in video synthesis, these metrics are similarly employed to measure view consistency across generated images under different viewpoints in this work.

First of all, LPIPS [20] is a widely used metric for evaluating the perceptual similarity between two images. It computes similarity by passing each image through a VGG network [12], extracting feature representations from intermediate layers, and comparing them. Mathematically, it is expressed as follows:

$$\text{LPIPS}(I_1, I_2) = \sum_l w_l \left\| \phi_l(I_1) - \phi_l(I_2) \right\|_2^2 \tag{1}$$

where: I_1, I_2 are the input images, $\phi_l(I)$ denotes the deep feature representation from layer l of a pre-trained network, w_l is a learned weight for layer l, $\| \cdot \|_2$ represents the Euclidean norm. And, CLIP Similarity [4] is a metric that measures the similarity between two images using the CLIP model. Specifically, it evaluates how similar the output embeddings are after passing the images through CLIP's encoder.

$$\text{CLIPScore}(I_1, I_2) = \frac{\psi(I_1) \cdot \psi(I_2)}{\|\psi(I_1)\| \|\psi(I_2)\|} \tag{2}$$

where: $\psi(I)$ represents the CLIP embedding of image I, $\cdot$ denotes the dot product, $\| \cdot \|$ represents the Euclidean norm. As illustrated in Fig. 2, LPIPS-*first* and CLIPSIM-*first* calculate LPIPS and CLIP similarity between the input viewpoint and all other viewpoints. In contrast, LPIPS-*next* and CLIPSIM-*next* measure consistency between adjacent viewpoints. Since the Next metrics are less affected by viewpoint changes compared to the First metrics, they provide a more effective measure of view consistency.

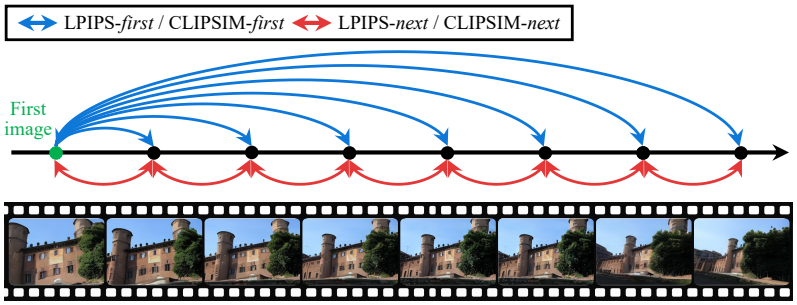

Figure 2. **Illustration of consistency metrics.** A figure that illustrates the primary consistency metrics used in this study: LPIPS-*first*, CLIPSIM-*first*, and LPIPS-*next*, CLIPSIM-*next* [3, 23]. The *first* metric measures LPIPS and CLIP similarity by comparing images from different viewpoints against the input viewpoint's image. In contrast, the *next* metric computes LPIPS and CLIP similarity by comparing images from adjacent viewpoints.

C.2. Camera parameter accuracy

View consistency is evaluated using the metrics outlined in Section C.1. However, these metrics are inherently limited in providing a comprehensive assessment. Consequently, additional evaluation is required for a more thorough analysis. Specifically, we propose estimating the camera viewpoint of the generated novel view synthesis images and comparing it against the ground truth camera parameters. This approach captures consistency aspects that previous metrics fail to address. To quantify camera parameter accuracy, three evaluation metrics are employed: (1) Frobenius Norm, (2) Rotation Angle Difference, and (3) Angular Consistency.

Frobenius Norm Frobenius Norm is a matrix norm that extends the Euclidean norm to matrices. It is defined as the square root of the sum of the absolute squares of the elements of a matrix. This norm provides a measure of the overall size of a matrix. Since the camera extrinsic parameter is a matrix composed of 3x4 or 4x4, it could provide the camera parameter accuracy between the estimated camera parameter and the ground truth camera parameter. Given the difference $|a_{ij}|$ between the measured camera parameters and the ground-truth camera parameters, the Frobenius Norm $\|A\|_F$ is calculated as:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}. \tag{3}$$

Rotation Angle Difference Rotation Angle Difference measures the angular discrepancy between two rotation matrices or quaternions. It is commonly used in 3D vision, robotics, and computer graphics to quantify rotational errors. Given two rotation matrices $R_1, R_2 \in SO(3)$, the Rotation Angle Difference θ is computed as:

$$\theta = \cos^{-1}\left(\frac{\text{trace}(R_1^T R_2) - 1}{2}\right). \tag{4}$$

Angular Consistency Angular Consistency refers to the property that ensures the relative orientation between viewpoints remains stable across transformations. It is particularly important in applications such as novel view synthesis, 3D reconstruction, and camera pose estimation. To quantify angular consistency, given a set of rotation matrices R_i associated with different viewpoints, the angular deviation between consecutive viewpoints can be measured as:

$$\theta_i = \cos^{-1}\left(\frac{\text{trace}(R_i^T R_{i+1}) - 1}{2}\right). \tag{5}$$

For a sequence of rotations, the overall angular consistency error can be defined as the variance of the angular differences where $\bar{\theta}$ is the mean rotation angle difference:

$$E_{\text{angular}} = \frac{1}{N-1}\sum_{i=1}^{N-1}(\theta_i - \bar{\theta})^2, \quad \bar{\theta} = \frac{1}{N-1}\sum_{i=1}^{N-1}\theta_i. \tag{6}$$

Ensuring low angular consistency error is critical for maintaining coherence in generated views and preventing distortions in viewpoint transitions.

D. Experiment Details

This section outlines the experimental setup, and additional experiments. We begin by describing the implementation details (Section D.1) and then introduce supplementary experiments conducted to further validate the proposed method. The experiments are divided into four main parts.

First, the primary experiment is described, focusing on the evaluation of the consistency metric and camera accuracy, as detailed in Section D.2. Next, the target-view image reconstruction metric experiments conducted on the MegaScenes dataset [15], where images are generated for specific viewpoints rather than continuous ones, are presented in Section D.3. We then detail image reconstruction experiments on the RealEstate10K(RE10K) sequence dataset in Section D.4. Finally, the methodology for extracting camera poses and performing the 3D Gaussian Splatting rendering task is explained in Section D.5.

D.1. Implementation details

For the novel view synthesis diffusion model, this work utilizes pre-trained models such as ZeroNVS [9] and MegaScenes [15], which generate images at a resolution of 256×256, maintained across all experiments. DDIM with 50 inference steps is adopted for sampling. Additionally, the attention map dropout technique from Tewel et al. [14] is implemented, using a fixed dropout ratio of 0.2, with no significant performance changes observed when varying the dropout ratio. For noise initialization, a Gaussian low-pass filter is applied, following the filter used in prior studies [8, 18]. During the noise application to latent variables, the noise level is fixed at 950 after the latent variables pass through the VAE encoder. The warp algorithm is implemented using the Pyrender library, equal to MegaScenes. Camera parameters are also configured in the OpenGL environment, as required by Pyrender. For warping, the depth map is extracted using the Depth Anything model [19], a monocular depth estimation model, following the approach of the previous work, MegaScenes.

D.2. Consistency and camera accuracy experiment

The experiments are conducted using the MegaScenes, RE10K, and DTU datasets. Mip-NeRF 360 dataset is excluded due to its limited number of scenes. The evaluation focuses on assessing image viewpoint consistency across consecutive viewpoints. These viewpoints are arranged in an *orbit pose*, with the input image set as the central viewpoint and surrounding poses forming an orbit configuration. The variation in *orbit pose* is illustrated as a gray line in Fig. 3, and in this study, we define the orbit pose with 19 viewpoints, fixing the radius at 1 and setting the rotation angle to 30 degrees. Thus, this experiment evaluates how well models generate consistent images when given a single input image and an orbit pose as input.

Camera parameters are generally divided into intrinsic and extrinsic parameters, with extrinsic parameters primarily used for camera pose evaluation. Extrinsic parameters consist of a rotation matrix and a translation matrix. However, since the translation matrix is highly sensitive to scale, our evaluation process focuses on the rotation matrix. To evaluate camera accuracy, the generated images are saved from the previous experiment and COLMAP [10, 11] is used to extract camera poses. COLMAP is chosen because it remains widely used for constructing camera parameter datasets [6, 7]. Since images are generated using an orbit pose, we use the corresponding ground truth camera parameters for evaluation.

While most cases in camera parameter estimation are suitable for evaluation, some instances result in failure cases. During the camera parameter estimation process for image sets, there are cases where the expected 19 camera parameters in the orbit pose setup are not obtained. This issue arises when COLMAP's SIFT algorithm fails to find correspondences between images, leading to missing camera parameters. Since the inability to establish correspondences indicates a lack of image consistency, we apply a strong penalty when evaluating camera accuracy. For example, when comparing the estimated camera parameters to the ground-truth orbit pose, cases, where only 2 out of 19 cameras are reconstructed, are handled by duplicating the available parameters to match the required 19 viewpoints before evaluation. This ensures that inconsistent images, which fail to establish correspondences, receive a penalty. However, simply duplicating poses in this manner may not be suitable for all scenarios, as the criteria for duplication vary depending on the number of available poses. It would be possible to apply the maximum error to camera poses that are not successfully estimated. For Frobenius Norm, the maximum error can be defined by comparing it with the identity matrix. Similarly, for Rotation Angle Difference and Angular Consistency, the maximum error is typically set to π (180° radians). Leveraging these predefined values, we can introduce penalty-based adjustments, which can be further explored in future evaluations.

D.3. Target view image reconstruction experiment

The target view image reconstruction experiment differs from the evaluation of consistency between consecutive images. This experiment involves generating paired images for different viewpoints, where the model generates a corresponding image from another viewpoint given an image from one viewpoint. Unlike the previous experiment, which measures viewpoint consistency, this experiment evaluates the model's ability to accurately generate images for specific viewpoints. We conduct

Table 1. **Target view image generation evaluation.** Quantitative evaluation of our method across diverse datasets (MegaScenes, DTU, RE10K, and Mip-NeRF 360). We report PSNR, SSIM, and LPIPS to evaluate reconstruction quality. To assess the consistency of the warped image regions, we include Masked PSNR, Masked SSIM, and Masked LPIPS [15]. Additionally, FID and KID are reported to measure the overall quality of image generation. Our approach improves performance over the baseline methods (MS, ZeroNVS) across overall metrics, demonstrating better view consistency and reconstruction fidelity.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Masked PSNR ↑	Masked SSIM ↑	Masked LPIPS ↓	FID ↓	KID ↓
MegaScenes								
ZeroNVS	7.69	0.150	0.611	11.09	0.653	0.268	46.05	0.029
ZeroNVS + WAVE	**10.70**	**0.335**	**0.497**	**18.14**	**0.823**	**0.160**	**32.93**	**0.016**
MegaScenes	12.28	0.432	0.395	23.36	0.877	0.094	**13.55**	0.005
MegaScenes + WAVE	**12.49**	**0.438**	**0.392**	**24.496**	**0.878**	**0.092**	15.03	**0.005**
DTU								
ZeroNVS	8.94	0.213	0.644	18.05	0.627	0.286	63.27	0.016
ZeroNVS + WAVE	**11.11**	**0.347**	**0.587**	**19.81**	**0.706**	**0.246**	**57.65**	**0.008**
MegaScenes	10.88	**0.394**	0.503	20.36	0.722	**0.211**	34.96	0.007
MegaScenes + WAVE	**11.95**	0.386	**0.460**	**20.72**	**0.722**	0.212	**22.59**	**0.004**
RE10K								
ZeroNVS	12.27	0.263	0.520	20.25	0.653	0.251	8.05	0.002
ZeroNVS + WAVE	**12.33**	**0.269**	**0.514**	**20.33**	**0.657**	**0.248**	**6.15**	**0.001**
MegaScenes	11.62	**0.309**	**0.494**	19.98	0.224	0.669	15.81	0.007
MegaScenes + WAVE	**11.78**	0.261	0.516	**20.04**	**0.650**	**0.252**	**10.20**	**0.004**
Mip-NeRF 360								
ZeroNVS	11.00	0.123	0.675	23.89	0.768	0.209	**78.64**	0.016
ZeroNVS + WAVE	**11.48**	**0.137**	**0.664**	**24.32**	**0.779**	**0.201**	78.93	**0.014**
MegaScenes	11.90	**0.182**	0.566	25.76	0.809	0.155	58.87	**0.009**
MegaScenes + WAVE	**12.29**	0.181	**0.560**	**26.24**	**0.813**	0.155	**58.48**	0.011

this experiment to evaluate whether our method negatively impacts performance in this aspect. This experiment follows a similar approach to MegaScenes [15].

Reconstruction experiments are conducted on the MegaScenes [15], RE10K [22], DTU [1], and Mip-NeRF 360 [2] datasets. For the MegaScenes dataset, we follow the publicly available test code to ensure consistency with previous work [15]. For other datasets, custom test configurations are created due to the inaccessibility of the test code. Given the model's inherent limitations in generating images with significant viewpoint changes, paired datasets with closer viewpoint pairs are constructed for RE10K, DTU, and Mip-NeRF 360. While our primary focus is improving viewpoint consistency, our method also yields improvements in reconstruction and image metrics. Notably, metrics such as Masked PSNR, Masked LPIPS, and Masked SSIM [15] as for evaluating paired image consistency, shows improved performance in our experimental results. The results demonstrate that our approach maintains consistency not only between the generated images but also between the input view and the generated images. Additionally, the experiment shows that addressing the inherent factors of the diffusion models leads to improved image quality in generating novel view images from specific viewpoints.

D.4. RE10K sequence evaluation

RE10K [22] dataset differs from other datasets in that it is a sequence dataset composed of videos, where each frame contains corresponding image and camera viewpoint information. The dataset consists of videos, so evaluation is restricted to predefined viewpoints rather than arbitrary ones. However, since it can provide consecutive viewpoint ground truth images unlike other datasets, we conduct this experiment to validate our method by evaluating the generated images using image generation and reconstruction metrics. For evaluation, a random input view image is selected from a video sequence, and 10 frames are skipped between successive viewpoints to generate a total of 6 target viewpoints. The generated images are then compared with their corresponding ground truth frames using LPIPS, PSNR, FID, and KID as evaluation metrics.

D.5. Downstream task

Recent studies have increasingly explored the application of novel-view diffusion models to 3D rendering tasks [9, 16, 17], leveraging the generative capabilities of these models in combination with 3D models for rendering purposes. Rather than solely focusing on novel-view synthesis, this work also investigates the generation of consistent images and evaluates their effectiveness in 3D rendering tasks. First, camera poses are extracted from generated image sets using COLMAP [10, 11].

Table 2. **Additional 3D rendering downstream tasks.** We present the additional quantitative results of experiments from different datasets where 3D Gaussian Splatting [5] performs 3D rendering by using those generated by our method and baseline.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
DTU			
ZeroNVS	23.53	0.847	0.127
ZeroNVS + WAVE	**25.56**	**0.871**	**0.101**
MegaScenes	20.43	0.809	0.142
MegaScenes + WAVE	**26.95**	**0.895**	**0.078**
RE10K			
ZeroNVS	24.74	0.883	0.107
ZeroNVS + WAVE	**26.48**	**0.900**	**0.088**
MegaScenes	21.23	0.817	0.145
MegaScenes + WAVE	**24.51**	**0.876**	**0.086**

The measured poses and generated images are then used as input to the 3D Gaussian splatting model [5] for rendering, with 3,000 training iterations. In addition, to mitigate potential errors in COLMAP's camera parameter estimation, cases where the number of camera poses is significantly lower than the number of images are excluded from evaluation. Due to space limitations in the main paper, results from additional datasets that couldn't be included are provided in Table 2.

D.6. Quantitative evaluation of adaptive warp-range selection effect

Conventional batch self-attention [14, 21, 23] aggregates all key-value pairs to compute the final representation. In contrast, our proposed warp-guided adaptive attention selectively determines the relevant viewpoint range required to generate a specific viewpoint by the adaptive warp-range selection. It then aggregates only the necessary key-vale pairs based on viewpoint changes. As shown in the main paper, applying the conventional batch self-attention method, widely used in previous studies, results in reduced camera accuracy. However, since ground-truth images are unavailable, evaluating the generated results alone does not provide a fully reliable assessment. To address this, we conduct a quantitative evaluation of camera accuracy and extend our experiments beyond the MegaScenes dataset to DTU and RE10K datasets. As presented in Table 3, our warp-guided adaptive attention significantly outperforms conventional batch self-attention. This result validates our hypothesis that the reference range should dynamically adjust according to viewpoint changes.

Table 3. **Camera accuracy in warp-guided adaptive attention.** We show quantitative results to evaluate whether warp-guided adaptive attention improves viewpoint accuracy. Our method, warp-guided adaptive attention, achieves higher camera pose accuracy compared to batch self-attention, which has been used in previous studies.

	Frobenius Norm (Rotation) ↓	Rotation Angle Difference ↓	Angular Consistency ↓
MegaScenes			
MegaScenes + *conventional batch attention*	0.412	0.299	17.13
MegaScenes + WAVE	**0.382**	**0.277**	**15.91**
DTU			
MegaScenes + *conventional batch attention*	0.383	0.281	16.09
MegaScenes + WAVE	**0.155**	**0.110**	**6.32**
RE10K			
MegaScenes + *conventional batch attention*	0.321	0.231	13.25
MegaScenes + WAVE	**0.149**	**0.108**	**6.20**

E. More Qualitative Results

E.1. Camera parameter visualization

We present a visualization of the camera parameters measured in the camera accuracy experiment. While camera accuracy is evaluated using only the rotation matrix from the camera extrinsic parameters, the translation matrix is visualized to provide

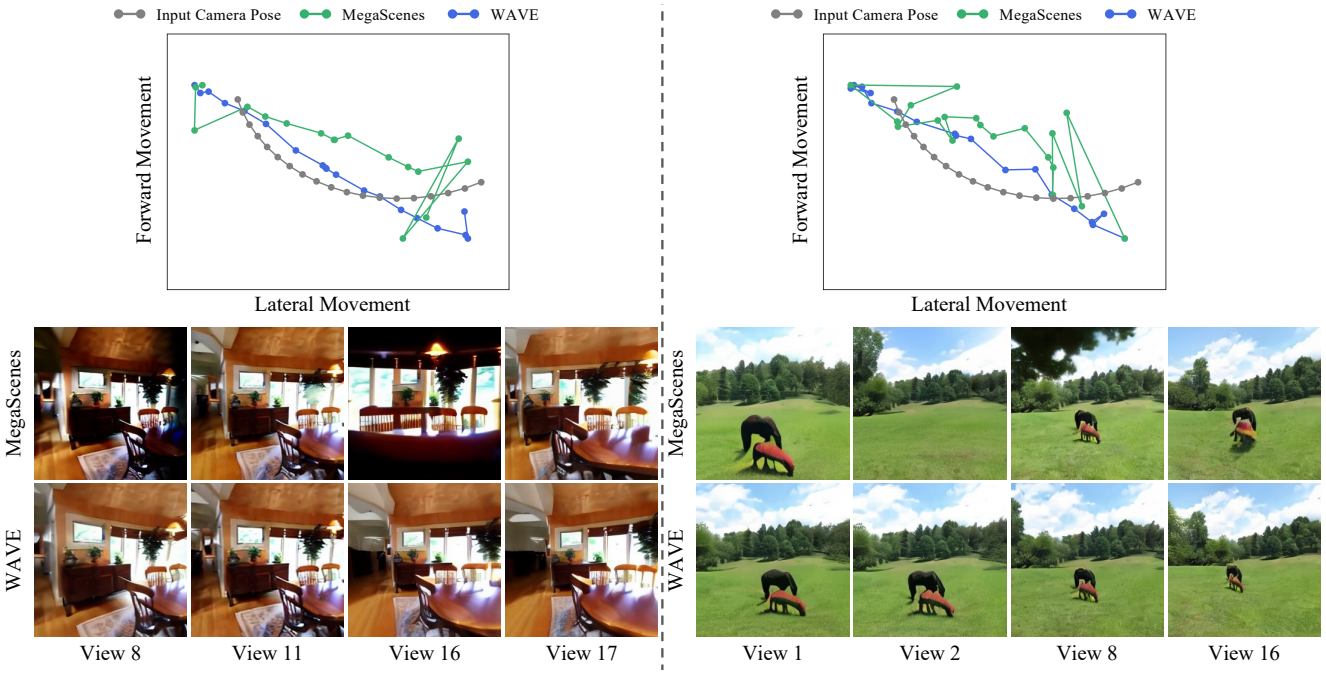

Figure 3. **Camera parameter visualization.** These examples are taken from the camera accuracy evaluation process. The images below represent the results generated by each model: MegaScenes [15] and MegaScenes + WAVE, while the camera graph above visualizes the measured camera poses from the generated images.

further insights into the translation component. Since the camera poses are normalized, the height variation is minimal. Therefore, only the x and z coordinates for camera translation are used. To maintain consistency with the ground-truth pose, the measured parameters are also normalized to match the corresponding scale. As illustrated in the Fig. 3, it shows that images with higher view consistency exhibit greater camera accuracy, whereas inconsistent images tend to have lower camera accuracy. This result demonstrates that evaluating camera parameters is an effective approach for assessing view consistency.

E.2. Additional samples

In Fig. 4, and Fig. 5, we provide additional generation results for diffusion-based methods. Fig. 4 presents results generated using the MegaScenes [15] dataset, while Fig. 5 showcases results from the RE10K [22] dataset. As previously noted in prior research [15], ZeroNVS [9] fails to properly reflect viewpoint changes, often producing artifacts and inconsistencies across images. In contrast, our method generates more consistent images, where objects remain persistently visible across different viewpoints, and color variations are minimized, compared to MegaScenes and ZeroNVS. Furthermore, to provide additional qualitative results, we present more examples from the RE10K and MegaScenes datasets in Fig. 6, and Fig. 7. These results demonstrate the generalizability of our method across multiple datasets.

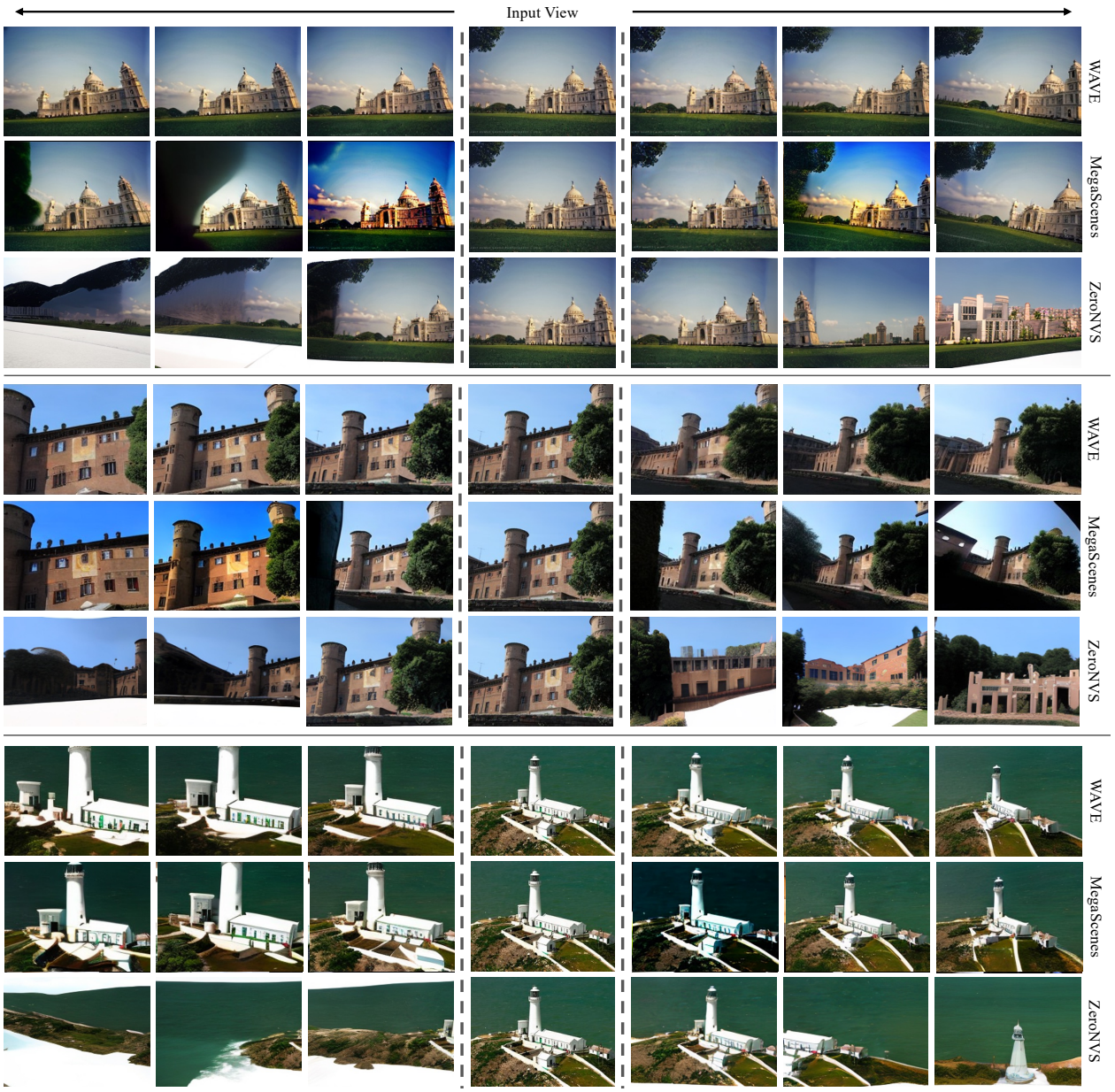

Figure 4. **Comparison to diffusion methods on MegaScenes.** We compare our framework with existing diffusion-based models, ZeroNVS [9] and MegaScenes [15]. Additional generation results are provided on the MegaScenes dataset. The images in the middle column represent the input images.

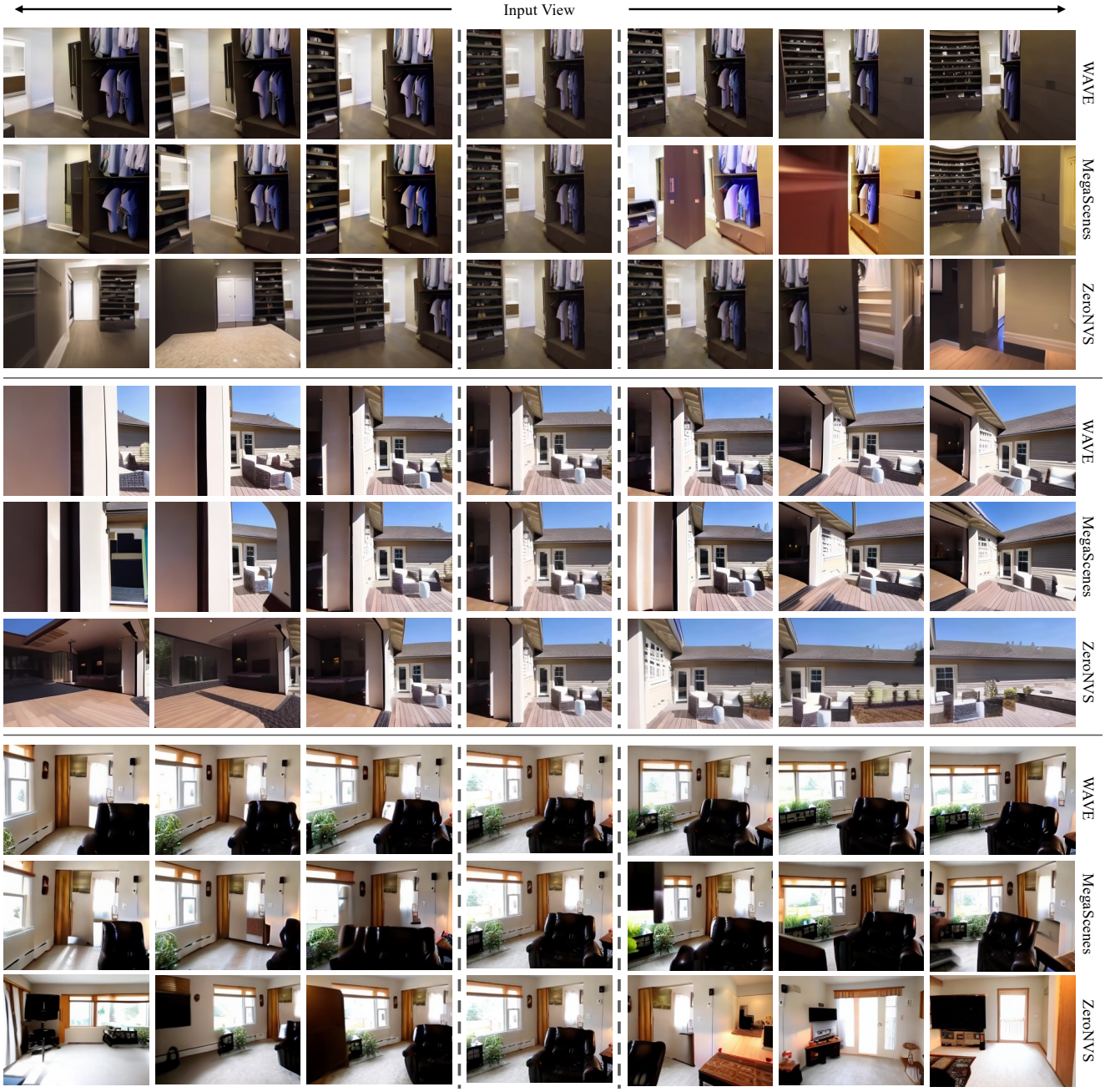

Figure 5. **Comparison to diffusion methods on RE10K.** We provide additional generation results on the RE10K [22] dataset to compare our method with baselines. The images in the middle column represent the input images.

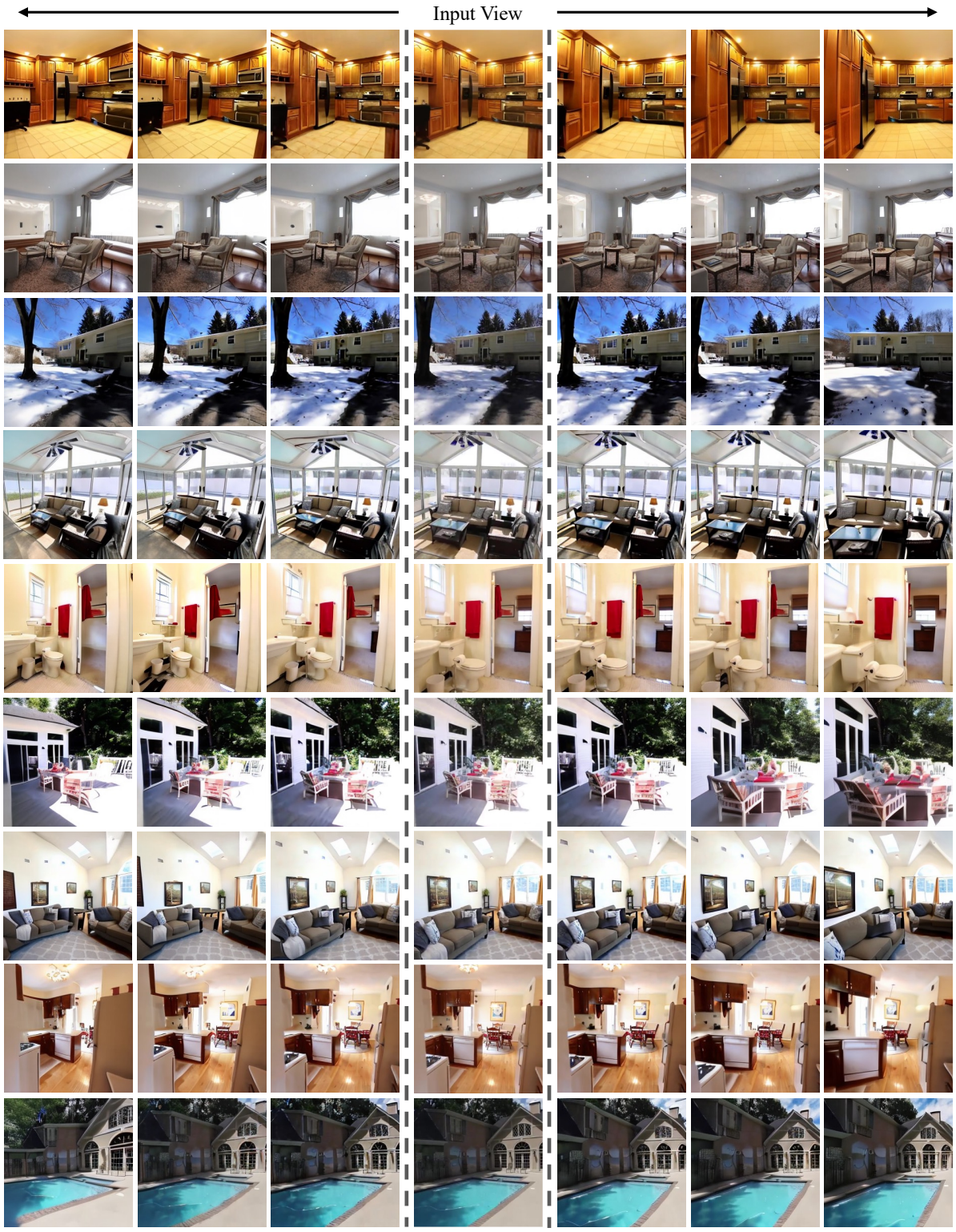

Figure 6. **Qualitative results.** Additionally, using samples from the RE10K dataset [22], we generate images with our method by taking the input view image in the middle column and continuous camera poses as inputs.

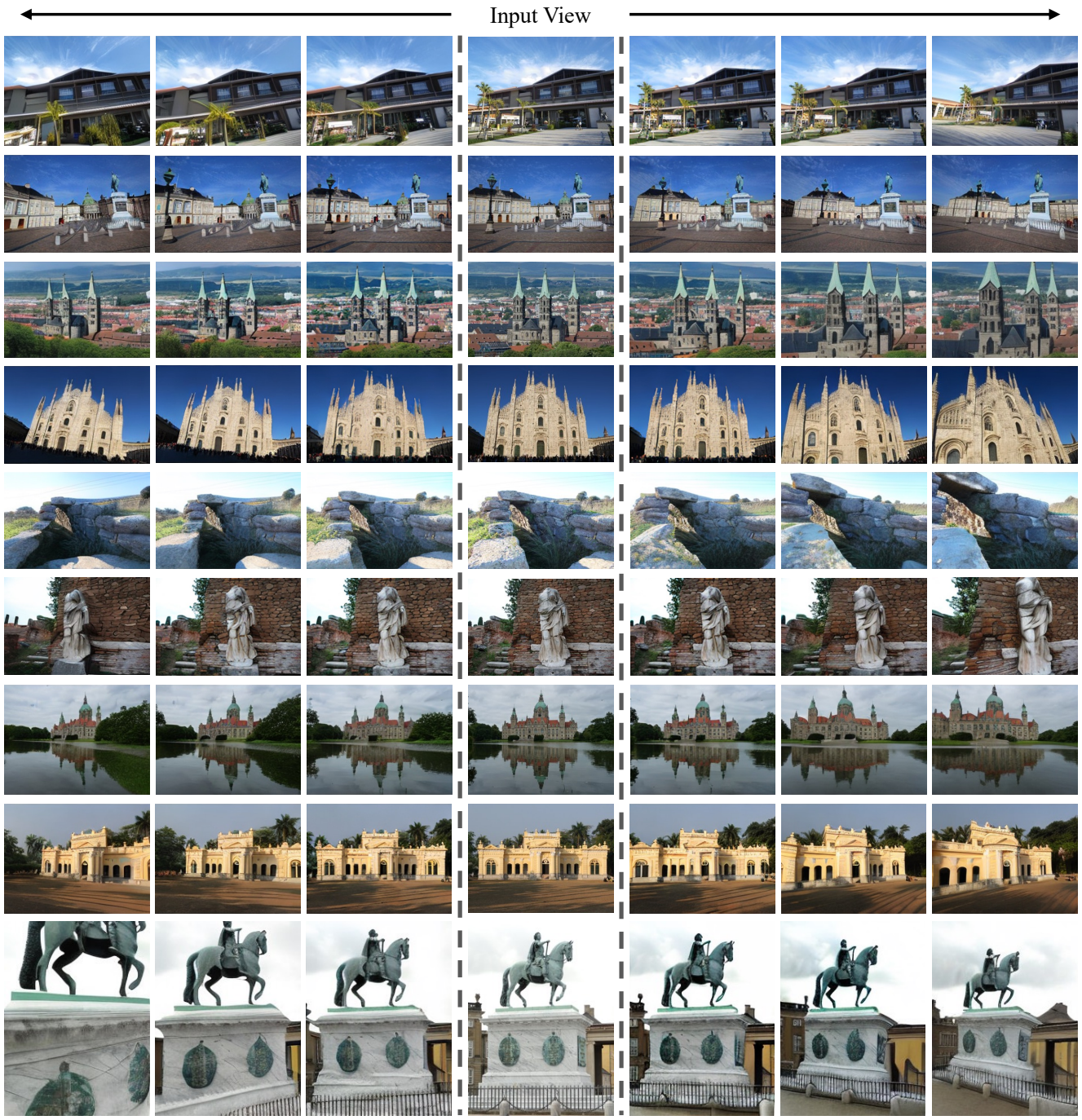

Figure 7. **Qualitative results.** Additionally, we generate images using our method with samples from the MegaScenes [15] dataset, where the input view image in the middle column and continuous camera poses serve as inputs.

F. Failure Cases

Our method relies on novel view synthesis diffusion models that generate images from specific viewpoints. Consequently, any inherent limitations of these models are reflected in the generated results. Additionally, since our approach incorporates 3D warping, it becomes challenging to extract meaningful information as the viewpoint difference increases. In Fig. 8, we present examples of two key issues: (1) diffusion models exhibit distortions, particularly in thin structures such as lines, and (2) as the viewpoint difference increases, the generated results exhibit repetitive structures, leading to reduced diversity. This phenomenon can be attributed to the diminishing information provided by the warped images as the viewpoint increases. However, as discussed in the main paper, the problem introduced by the warping algorithm could be alleviated through an autoregressive approach, where a specific range is first generated and then iteratively used to synthesize subsequent ranges. This approach could serve as a future direction for overcoming the limitations of the warping algorithm.

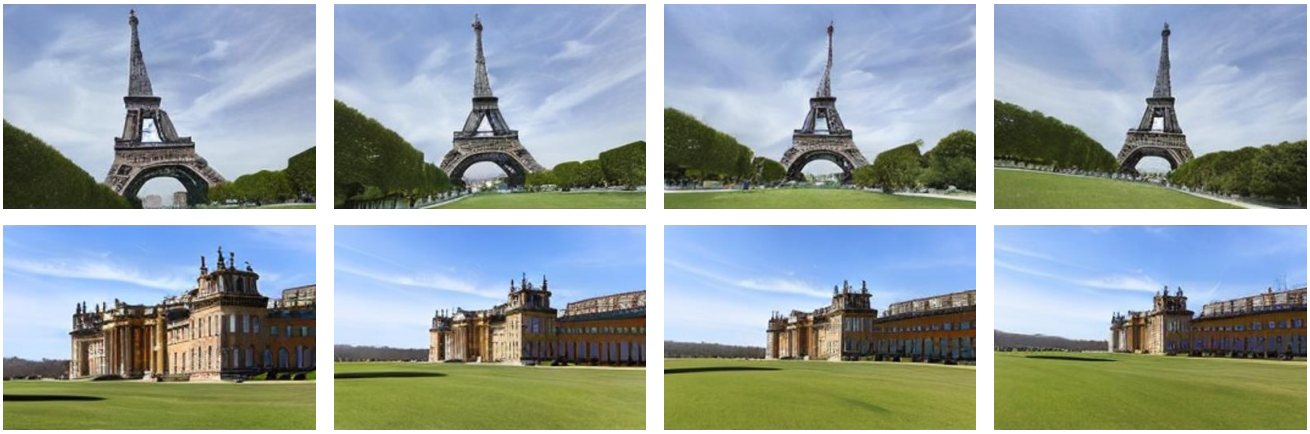

Figure 8. **Failure cases.** *Top*: an example that shows that the model struggles to accurately generate objects with thin structures, such as lines. *Bottom*: an example that shows the model's tendency to repeatedly generate identical structural patterns as the viewpoint difference increases, reducing diversity.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, pages 1–16, 2016. 5

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 5

[3] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer, 2024. 2, 3

[4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 2

[5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 6

[6] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 4

[7] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 4

[8] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *Transactions on Machine Learning Research*. 4

[9] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9420–9429, 2024. 4, 5, 7, 8

[10] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 5

[11] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 4, 5

[12] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[13] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 2

[14] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024. 4, 6

[15] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. In *European Conference on Computer Vision*, pages 197–214. Springer, 2025. 4, 5, 7, 8, 11

[16] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2024. 5

[17] Haiping Wang, Yuan Liu, Ziwei Liu, Wenping Wang, Zhen Dong, and Bisheng Yang. Vistadream: Sampling multiview consistent images for single-view scene reconstruction. *arXiv preprint arXiv:2410.16892*, 2024. 5

[18] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. In *European Conference on Computer Vision*, pages 378–394. Springer, 2025. 4

[19] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 4

[20] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2

[21] Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9720–9731, 2024. 6

[22] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 5, 7, 9, 10

[23] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024. 2, 3, 6