WAVE: Warp-Based View Guidance for Consistent Novel View Synthesis Using a Single Image

Jiwoo Park Tae Eun Choi Youngjun Jun Seong Jae Hwang*

Yonsei University

{wldn1677, teunchoi, youngjun, seongjae}@yonsei.ac.kr

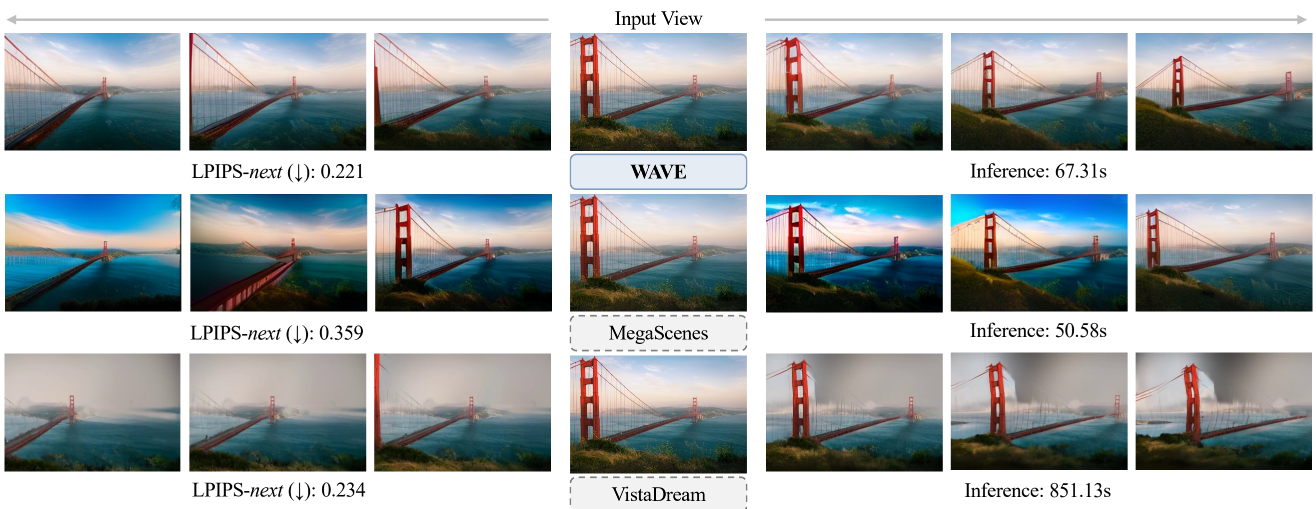

Figure 1. **Teaser.** We present WAVE, a training-free approach that enhances view consistency both among generated images and between the generated images and the input view. This approach is designed for novel view synthesis diffusion models using a input view. WAVE generates consistent images (lowest LPIPS-*next*) across views without extra modules, enabling roughly 67s of inference time for 19 images.

Abstract

Generating high-quality novel views of a scene from a single image requires maintaining structural coherence across different views, referred to as view consistency. While diffusion models have driven advancements in novel view synthesis, they still struggle to preserve spatial continuity across views. Diffusion models have been combined with 3D models to address the issue, but such approaches lack efficiency due to their complex multi-step pipelines. This paper proposes a novel view-consistent image generation method which utilizes diffusion models without additional modules. Our key idea is to enhance diffusion models with a training-free method that enables adaptive attention manipulation and noise reinitialization by leveraging view-guided warping to ensure view consistency. Through our comprehensive metric framework suitable for novel-view datasets, we show that our method improves view consistency across various diffusion models, demonstrating its broader applicability [1]

[1] *Project page:* jwoo-park0.github.io/wave

1. Introduction

Scene-level novel view synthesis from a single image has received significant interest for its potential to generate unseen views from only one image [5, 13, 35, 38]. In single-image scenarios where multi-view information is unavailable, generative models have become a key component in novel view synthesis compared to rendering-based models [10, 19]. Specifically, diffusion models with strong generative capabilities have been adopted to generate novel views by utilizing view prior knowledge [6, 30]. Such approaches have led studies to explore diffusion models trained on large-scale datasets of images and camera parameters to generate novel views from a single image [25, 32]. Leveraging their generative power, diffusion models can synthesize unseen scenes with zero-shot capability.

While diffusion-based models such as ZeroNVS [25] and MegaScenes [32] have demonstrated strong performance, a key limitation of these models is that they cannot maintain view consistency among generated images from different viewpoints, as shown in Fig. 2. For example, MegaScenes shows view inconsistencies due to variations in object appearance and color saturation, as illustrated in the second row of Fig. 1. To achieve view consistency, other studies [29, 35] have explored integrating diffusion models with additional models such as 3D models and large language models. In particular, VistaDream [35] leverages the 3D structures and contextual knowledge of vision language models to enhance consistency. Nevertheless, their applicability remains limited due to high computational costs and the complexity of multi-step pipelines. As shown in the third row of Fig. 1, despite the inference time being nearly ten times longer than that of MegaScenes, VistaDream still struggles to achieve satisfactory image quality.

Even with various efforts to complement diffusion models, inherent factors within diffusion models that compromise view consistency remain unresolved [11, 20]. A major factor is that existing diffusion models generate each image separately [25, 32], disregarding the relationship between images. As a result, generated images lack interdependencies, making it challenging to maintain view consistency. Additionally, diffusion models exhibit high sensitivity to noise randomness during the generation process [8, 24, 37], leading to notable variations in the generated outputs. This variability impairs the model's ability to maintain overall visual coherence. Addressing these intrinsic factors directly could be a practical framework for achieving both view consistency and comparable inference time. Therefore, we introduce a streamlined method, WAVE, for view consistency that addresses the challenges inherent in diffusion models without the need for extra modules or training.

To strengthen image interdependencies, we propose **warp-guided adaptive attention**, which integrates 3D warping [18] into batch self-attention to generate all viewpoints simultaneously. Batch self-attention computes attention by aggregating different keys and values across the batch, enabling information sharing between images [21, 31, 41]. However, batch self-attention fails to account for viewpoint changes, limiting its ability to capture view-dependent information in novel view synthesis. Thus, we utilize 3D warping which shifts pixels of an image from the input view to an arbitrary view. Integrated as a prior in batch self-attention, 3D warping provides cues for viewpoint transformations, improving view consistency. Moreover, noise randomness introduces unintended variations in generated images, diminishing view consistency. We propose **pose-aware noise initialization** which embeds the target image information into the initial diffusion noise to reduce noise randomness. Providing low-frequency information to the initial noise of diffusion models has proven effective in mitigating randomness [21, 37]. Building on this, our method adopts the low-frequency of warped images to inject pose-aware information into the initial noise.

Figure 2. **View inconsistency in diffusion models.** Generated novel views using the single image on the left are presented. WAVE maintains view consistency, whereas other diffusion-based models like MegaScenes [32] and ZeroNVS [25] struggle to preserve consistency across viewpoints.

As warped images inherently retain partial low-frequency details of the target view, they help alleviate noise randomness, ultimately enhancing view consistency.

In a evaluation of view consistency improvement, we construct the comprehensive view consistency metric framework. This framework combines video metrics [7, 41] and camera pose accuracy metrics as reconstruction-based metrics are unreliable for consistency evaluation [12]. Additionally, we evaluate image quality on a sequential dataset and validate our method in a downstream task using 3D Gaussian splatting [10]. The contributions are as follows:

- We propose WAVE, a training-free method for scene-level novel view synthesis, adaptable to various diffusion models without additional training.
- We present two approaches to enhance view consistency: (1) warp-guided adaptive attention to facilitate inter-view information flow, and (2) pose-aware noise initialization to mitigate the negative impact of noise randomness.
- Through our comprehensive metric framework along with extensive qualitative results, WAVE achieves improved performance over diffusion-based models, demonstrating comparable results to the rendering model.

2. Related Work

2.1. Novel View Synthesis from Single Scene Image

Scene-level novel view synthesis from a single image has emerged as a crucial area of research, offering a feasible solution for generating realistic images on novel viewpoints while relying only on a single image. Unlike object-centric approaches [4, 15, 27, 34] that focus on generating isolated objects, scene-level synthesis [13, 14, 38] aims to generate complex and complete environments.

Recent advancements in this domain have leveraged diffusion-based models, as they possess the generative capacity to generate novel views [11, 14, 20]. Consequently, efforts have been made to develop large-scale trained models [25, 32] on Stable Diffusion [23]. ZeroNVS [25] introduces scene-scale variables for fine-tuning at scene level. MegaScenes [32] has since addressed the viewpoint inaccuracies and limited generalization ability of ZeroNVS by employing a warping strategy for scene images. Various works [11, 20] have also focused on enforcing consistency between the input view and the generated novel view images. Despite these contributions, ensuring robust consistency across multiple views synthesized by diffusion models remains a fundamental challenge.

To ensure view consistency, approaches such as VistaDream [35] and RealmDreamer [29] have integrated diffusion models with 3D models, reconstructing 3D from a single image or text inputs by leveraging 3D-aware representations. Although these approaches mitigate view inconsistency, they are made of multi-step pipelines that increases computational cost, making them less practical. While previous works have focused on auxiliary modules to assist diffusion models, the factors inherent to diffusion models which negatively impact view consistency remain unexplored. Accordingly, a practical approach that resolves these inherent factors needs to be explored.

2.2. Image Consistency in Diffusion Models

Diffusion models struggle to maintain consistency across image sequences in the fields of video generation [21, 41], text-to-image synthesis [8, 17, 31], and image synthesis [3, 36]. Various methods have been proposed for consistency by leveraging attention or cross-view conditioning without additional training [31, 39, 41]. ConsiStory [31] proposes a subject attention masking method to ensure consistent subject generation across different prompts, while StoryDiffusion [41] introduces an attention mechanism that enables the model to share contextual information.

Other image-to-video generation methods [9, 21, 37] have emphasized the negative effect of noise randomness on consistency, attributing this issue to the discrepancy between noise in the training and inference stages. FreeInit [37] introduces iterative sampling to mix low-frequency information from latent variables into the noise, while ConsistI2V [21] incorporates low-frequency information from the first frame into the initial noise to mitigate randomness.

These methods have primarily focused on preserving structural continuity. In novel view synthesis, however, attempts to consistently maintain view-dependent spatial information have yet to be made. Therefore, to ensure view consistency in novel view synthesis where view variation is a key factor, integrating geometric transformations for view changes into consistency methods is essential.

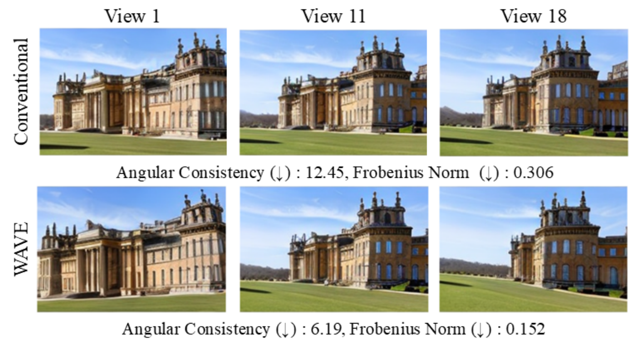

Figure 3. **Effectiveness of adaptive warp-range selection.** We present examples of applying our method alongside the conventional batch self-attention method [31, 39] that references all viewpoints. The differences in the two camera pose accuracy metrics, Angular Consistency and Frobenius Norm, show that the previous method generates images that are less accurately aligned with the camera poses than ours.

3. Methods

We propose a training-free method that leverages warped images for consistent scene-level novel view synthesis using diffusion models pre-trained on large-scale datasets with camera poses and images [25, 32].

3.1. Preliminaries

3D warping. 3D warping [18] is the process of transforming an input view image to match a desired camera pose using its depth. The transformation involves unprojecting image pixels into 3D space using depth values, followed by reprojecting them according to the specified camera poses. This approach enables the generation of warped images that accurately reflect the new viewpoints.

Batch self-attention. Self-attention [33] extracts image features by applying linear projections with the attention matrices W_q, W_k, W_v to produce Q, K, V. The attention map A is then computed as $\texttt{softmax}(QK^T/\sqrt{d_k})$ and multiplied by V to generate hidden features h. Otherwise, batch self-attention [31, 39] aggregates the keys and values from all N generations in a batch. Batch self-attention is defined as:

$$\begin{aligned}
&Q_i \in \mathbb{R}^{p\times d_q}, \quad K_i \in \mathbb{R}^{p\times d_k}, \quad V_i \in \mathbb{R}^{p\times d_v}, \\
&K^* = [K_1, K_2, \dots, K_N] \in \mathbb{R}^{N\cdot p\times d_k}, \\
&V^* = [V_1, V_2, \dots, V_N] \in \mathbb{R}^{N\cdot p\times d_v}, \\
&A_i = \texttt{softmax}(Q_i K^{*\top}/\sqrt{d_k}) \in \mathbb{R}^{p\times N\cdot p}, \\
&h_i = A_i \cdot V^* \in \mathbb{R}^{p\times d_v}. \qquad (1)
\end{aligned}$$

Batch self-attention enables the model to generate images from multiple viewpoints in a single forward pass, preventing diffusion models from generating images independently. However, referencing all views with non-overlapping and unnecessary information reduces the reliance on the desired camera pose view features. This makes

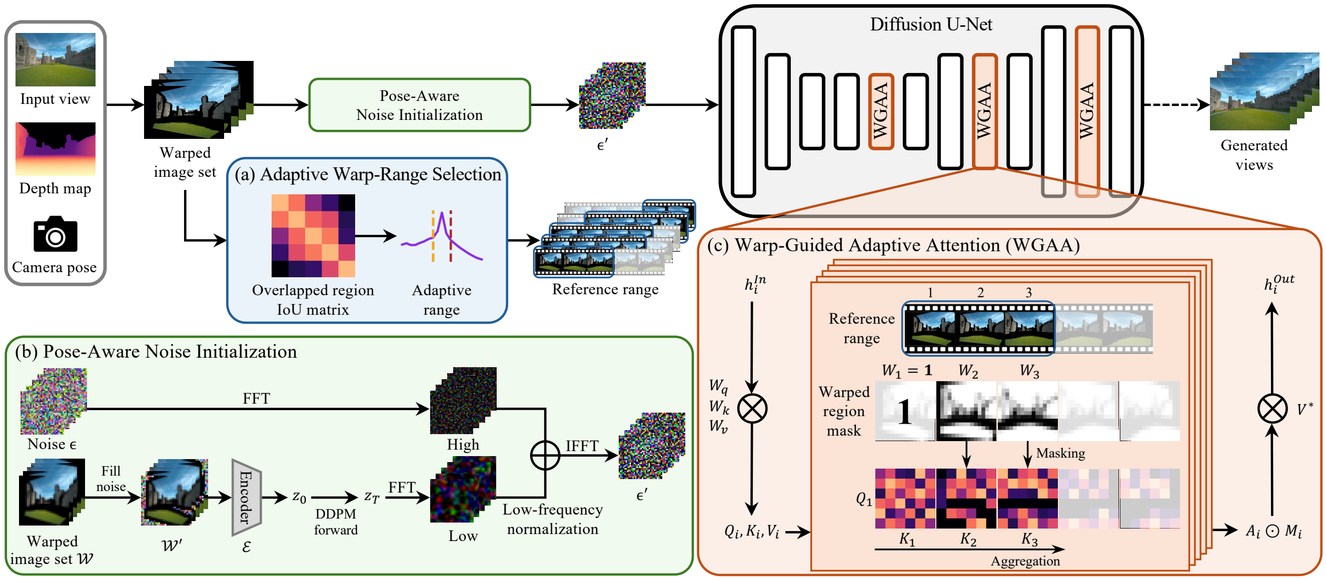

Figure 4. **Illustration of WAVE.** Given an input view, depth map and continuous camera poses, our method generates scene images with smooth view transitions through three distinct processes: (a) adaptive warp-range selection utilizes warped region masks, from which the relevance between viewpoints is determined to compute the reference range for attention. (b) pose-aware noise initialization (PANI) re-initializes the diffusion initial noise by leveraging warped images and initial noise, incorporating frequency domain information. (c) warp-guided adaptive attention (WGAA) utilizes the warped region masks and reference range obtained from the adaptive warp-range selection, performing masked batch self-attention.

it difficult to generate an image that matches the desired camera pose. To adequately reflect view variations, instead of referring to all views, we adaptively select the reference views (Section 3.2) and develop an attention mechanism that masks non-overlapping regions within these selected views (Section 3.3).

3.2. Adaptive Warp-range Selection

Referencing all other viewpoints in the batch as in batch self-attention [31, 39] reduces viewpoint accuracy, as shown in Fig. 3 and Supp. Table 3. Fig. 3 also shows examples of misalignment with the desired viewpoint. Aggregating information from non-overlapping views ultimately degrades viewpoint accuracy. Therefore, we set a reference range of viewpoints within the attention mechanism to utilize information from only a relevant subset of viewpoints.

To determine the optimal reference range, we perform the adaptive warp-range selection shown in Fig. 4a. We utilize warped images from 3D warping to compute viewpoint relevance. As the viewpoint shifts, regions that are not visible from the input view become missing regions in the warped images, assigned a value of zero. In order to create warped region masks M, we convert warped images into binary regions $W_{1:N}$, where pixels with a value of zero remain 0, and non-zero pixels are set to 1. Then, to measure the relevance between viewpoints (*e.g.*, i and j), we compute the IoU u_{ij} on warped region masks to construct the matrix $U = \{u_{ij}\} \in \mathbb{R}^{N\times N}$, which provides a precise measure of their spatial relevance [22]. Based on this matrix, the adaptive range $[\mu_i - \sigma_i,\ \mu_i + \sigma_i]$ is selected for each viewpoint, for mean μ_i and standard deviation σ_i. Applying the standard deviation of the IoU matrix enables the dynamic determination of reference points, making our approach applicable across various sets of viewpoints.

3.3. Warp-Guided Adaptive Attention

Moreover, in batch self-attention, as the number of referenced images increases, each image becomes overly dependent on information from other images rather than its own. To prevent excessive reliance on external information, previous work proposed a method of subject attention masking based on the subject mask to facilitate the selection of information [31]. Attention masking allows the model to selectively focus on the important information. Building on this insight, we design warp-guided attention masking, which employs warped region masks in the diffusion decoder as shown in Fig. 4c, as follows:

$$\begin{aligned} A_i &= \texttt{softmax}(Q_i K^{*\top}/\sqrt{d_k}), \\ W_{1:N} &= \mathcal{T}(\mathcal{I},\, \texttt{pose}_{1:N}) \odot \mathbf{1}_{\mathcal{T}(\mathcal{I},\, \texttt{pose}_{1:N})\neq 0}, \\ M_i &= [W_1, W_2, \ldots, \mathbf{1}, \ldots, W_N], \\ h_i &= (A_i \odot M_i)\cdot V^* \in \mathbb{R}^{p\times d_v}, \end{aligned} \tag{2}$$

where $\mathcal{T}$ is a 3D warping operation. This approach leverages the assumption that the decoder's attention mechanism encodes spatial information relevant to different views. For details of the assumption, please refer to the supplementary

(a) Without noise (b) Fill noise (c) WAVE

Figure 5. **Analysis of warping methods.** We present the generation results of the methods attempted to mitigate the negative impact of missing regions in warped images : (a) not filling missing regions with noise, (b) only filling missing regions with noise, (c) our method (low-frequency normalization and filling noise).

material. These warped region masks encapsulate the viewpoint cues of the target view image to be generated. The warped region masks effectively serve as a masking technique by capturing the spatial differences between viewpoints, allowing for a focused integration of important information. Warp-guided adaptive attention performs warped region mask attention based on the adaptive range obtained through adaptive warp-range selection. This enables the attention mechanism to effectively reflect viewpoint changes.

3.4. Pose-Aware Noise Initialization

Beyond independent generation, noise randomness in diffusion models is another key issue. The stochastic nature of noise adversely impacts image consistency [8, 24, 37], which extends similarly to view consistency.

View consistency, unlike image consistency, requires integrating information regarding view variations. Therefore, we propose pose-aware noise initialization which incorporates such information for noise reinitialization using warped images to enhance view consistency, as illustrated in Fig. 4b. Warped images capture the partial low-frequency of the target images, reflecting variations across different viewpoints. Leveraging them for noise reinitialization mitigates noise randomness by integrating geometric information from view changes into the noise. The process is summarized in Algorithm 1. Warped images $\mathcal{W}$ are generated, aligned with the given camera poses. These images are passed through the VAE encoder $\mathcal{E}$, and DDPM noise is then added to make z_T. Low-frequency information is extracted from z_T and mixed with the high-frequency components of the noise to preserve essential information in warped images while introducing randomness in the high-frequency parts to enhance visual details [21, 37].

Warped images contain missing regions due to the absence of information from unseen viewpoints. Injecting low-frequency information directly from warped images results in undesirable black holes, as shown in Fig. 5a. This occurs because missing regions are reflected in the generated outputs. To address this, we apply random noise to fill the missing regions $\mathcal{W}'$ before injection, suppressing low-frequency associated with these regions. Otherwise, the low-frequency components of the random noise interfere with the original low-frequency of the warped images, leading to artifacts, as illustrated in Fig. 5b. To mitigate this, we normalize low-frequency components based on their magnitudes, controlling the impact of the random noise. This suppression of noise, as shown in Fig. 5c, ensures a more consistent generation process.

Algorithm 1 Pose-Aware Noise Initialization

Input: Reference image $\mathcal{I}$, camera poses $\texttt{pose}_{1:N}$,
Depth map $\mathcal{D}$
Output: re-initialized noise ϵ'
Note: Gaussian filter $\mathcal{G}(\mathcal{D}_0)$, VAE encoder $\mathcal{E}$

Apply 3D Warping Algorithm using $\mathcal{I}$, $\mathcal{D}$, and $\texttt{pose}_{1:N}$
$\mathcal{W} \leftarrow \{\mathcal{W}_1, \mathcal{W}_2, \dots, \mathcal{W}_n\}$
$\mathcal{W}' \leftarrow \mathcal{W} + \eta \cdot \mathbf{1}_{W=0}$ ▷ Add noise η to zero regions of W
$z_0 \leftarrow \mathcal{E}(W)$, Sample $\epsilon \sim \mathcal{N}(0, I)$
$z_T \leftarrow$ DDPM Forward(z_0, T) ▷ noise scheduler
$F_T^{\text{low}} \leftarrow$ FFT$(z_T) \odot \mathcal{G}(\mathcal{D}_0)$
$F_T^{\text{low}} \leftarrow F_T^{\text{low}}/\text{mean}(|F_T^{\text{low}}|)$ ▷ mean of F_T^{low}'s magnitude
$F_\epsilon^{\text{high}} \leftarrow$ FFT$(\epsilon) \odot (1 - \mathcal{G}(\mathcal{D}_0))$
$\epsilon' \leftarrow$ IFFT$(F_T^{\text{low}} + F_\epsilon^{\text{high}})$
Return ϵ'

4. Experiments

Baselines & Datasets. Our method is compared with other diffusion-based approaches, including MegaScenes [32] and ZeroNVS [25], as well as a multi-step pipelines model VistaDream [35]. Experiments conducted on the datasets MegaScenes [32], RealEstate10K (RE10K) [40], DTU [1], and Mip-NeRF 360 [2] encompass synthetic, indoor, and outdoor scenes to ensure our method's generalizability.

Metric framework. Several novel-view datasets [1, 32] lack sequential viewpoints and ground-truth images, complicating the use of reconstruction metrics for consistency evaluation. The reliability of reconstruction metrics in evaluating consistency has been questioned [12]. To address this, we design a view consistency metric framework that jointly evaluates image consistency and camera pose accuracy. This framework includes video metrics (LPIPS-*next* / *first* and CLIPSIM-*next* / *first*) alongside camera pose accuracy metrics (Frobenius Norm of rotation matrix, Rotation Angle Difference, and Angular Consistency). For further details, please refer to Supp. Section C.

Experimental results based on our view consistency metric framework, along with results for reconstruction quality using the video dataset [40] are presented in Section 4.1. Additionally, Section 4.2 presents qualitative comparisons against existing methods, Section 4.3 covers the ablation study, and Section 4.4 evaluates our method on a down-

Table 1. **Consistency and camera accuracy metrics.** We conduct experiments on the three datasets MegaScenes, DTU, and RE10K and evaluate consistency using two metrics: *Next* metric, which assesses consistency with neighboring viewpoints, and *First* metric, which measures consistency with the input view. Additionally, to evaluate camera accuracy, a crucial aspect of novel view synthesis, we report the accuracy of camera extrinsic parameters using Frobenius Norm, Rotation Angle Difference, and Angular Consistency.

Method	*Next*		*First*		Frobenius Norm (Rotation) ↓	Rotation Angle Difference ↓	Angular Consistency ↓
	LPIPS ↓	CLIPSIM ↑	LPIPS ↓	CLIPSIM ↑			
MegaScenes							
ZeroNVS	0.561	0.887	0.602	0.826	0.563	0.410	23.53
ZeroNVS + WAVE	**0.404**	**0.933**	**0.545**	**0.851**	**0.557**	**0.404**	**23.17**
MegaScenes	0.607	0.861	0.609	0.813	0.481	0.355	20.35
MegaScenes + WAVE	**0.397**	**0.920**	**0.561**	**0.826**	**0.382**	**0.277**	**15.91**
DTU							
ZeroNVS	0.489	0.883	0.726	0.717	0.568	0.414	23.77
ZeroNVS + WAVE	**0.344**	**0.925**	**0.704**	**0.739**	**0.550**	**0.401**	**23.00**
MegaScenes	0.379	0.898	0.692	0.697	0.308	0.227	13.04
MegaScenes + WAVE	**0.130**	**0.956**	**0.634**	**0.716**	**0.155**	**0.110**	**6.32**
RE10K							
ZeroNVS	0.478	0.897	0.764	**0.745**	0.624	0.459	26.30
ZeroNVS + WAVE	**0.388**	**0.921**	**0.761**	0.743	**0.620**	**0.455**	**26.10**
MegaScenes	0.347	0.912	**0.722**	0.676	0.242	0.179	10.25
MegaScenes + WAVE	**0.262**	**0.948**	0.728	**0.695**	**0.149**	**0.108**	**6.20**

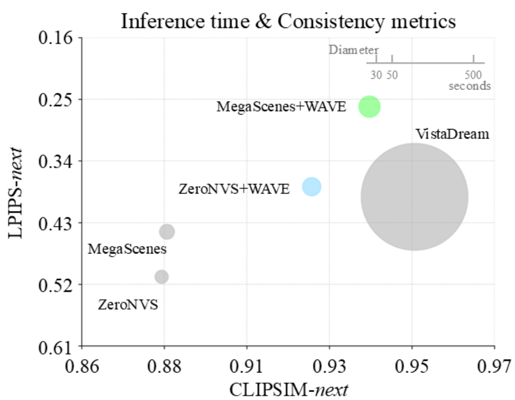

Figure 6. **Overall comparison of methods.** The size of the circles represents inference time. Applying our method improves performance with comparable computation costs.

stream task. Additional results and experimental settings can be found in the supplementary material.

4.1. Comparison with Baselines

Table 1 compares diffusion-based models with our method. Our method steadily outperforms diffusion-based models across diverse datasets, showing adaptability to diffusion models. Specifically, on the MegaScenes dataset, our method achieves 0.397 in LPIPS-*next* and 0.920 in CLIPSIM-*next*, while generating images with higher accuracy in camera viewpoints. VistaDream [35] incorporates a rendering method giving it an advantage in consistency metrics over diffusion models. Thus, VistaDream is compared based on consistency metrics *and* inference time, unlike other models. Notably, as shown in Fig. 6, the result of MegaScenes+WAVE exhibits significantly lower inference time than VistaDream while achieving comparable performance. This shows that solving the inherent factors of diffusion models can achieve a performance similar to rendering-based methods without excessive computational cost, demonstrating the practicality of our method.

Table 2. **RE10K sequence evaluation.** We present the results of image reconstruction and generation quality on RE10K, a video sequence dataset with varying viewpoints, demonstrating that our method outperforms the baseline.

Method	PSNR ↑	LPIPS ↓	FID ↓	KID ↓
VistaDream	10.88	0.567	29.48	0.011
ZeroNVS	10.84	0.612	24.13	0.010
ZeroNVS + WAVE	11.38	0.602	25.67	0.009
MegaScenes	12.02	0.469	18.20	0.006
MegaScenes + WAVE	**13.27**	**0.450**	**15.78**	**0.004**

Moreover, to evaluate image quality not captured by our metric framework, we employ the image sequence (*i.e.*, video) dataset, RE10K, to assess the reconstruction and generation capabilities (Table 2). This dataset contains both ground-truth images and camera poses for varying views, allowing the evaluation of image quality. The results show that our method outperforms existing models in both image generation (FID, KID) and reconstruction (PSNR, LPIPS) metrics. The results demonstrate that our method produces consistent images while achieving higher image quality.

4.2. Qualitative Comparison

We compare our method against baselines with qualitative results. As shown in Fig. 7, MegaScenes struggles to maintain consistency across different viewpoints. In contrast, WAVE produces more coherent results, addressing the limitations of existing diffusion models. While VistaDream maintains view consistency, it exhibits prominent artifacts in generated images. WAVE achieves superior performance compared to VistaDream, demonstrating that better-quality images can be generated without relying on multiple mod-

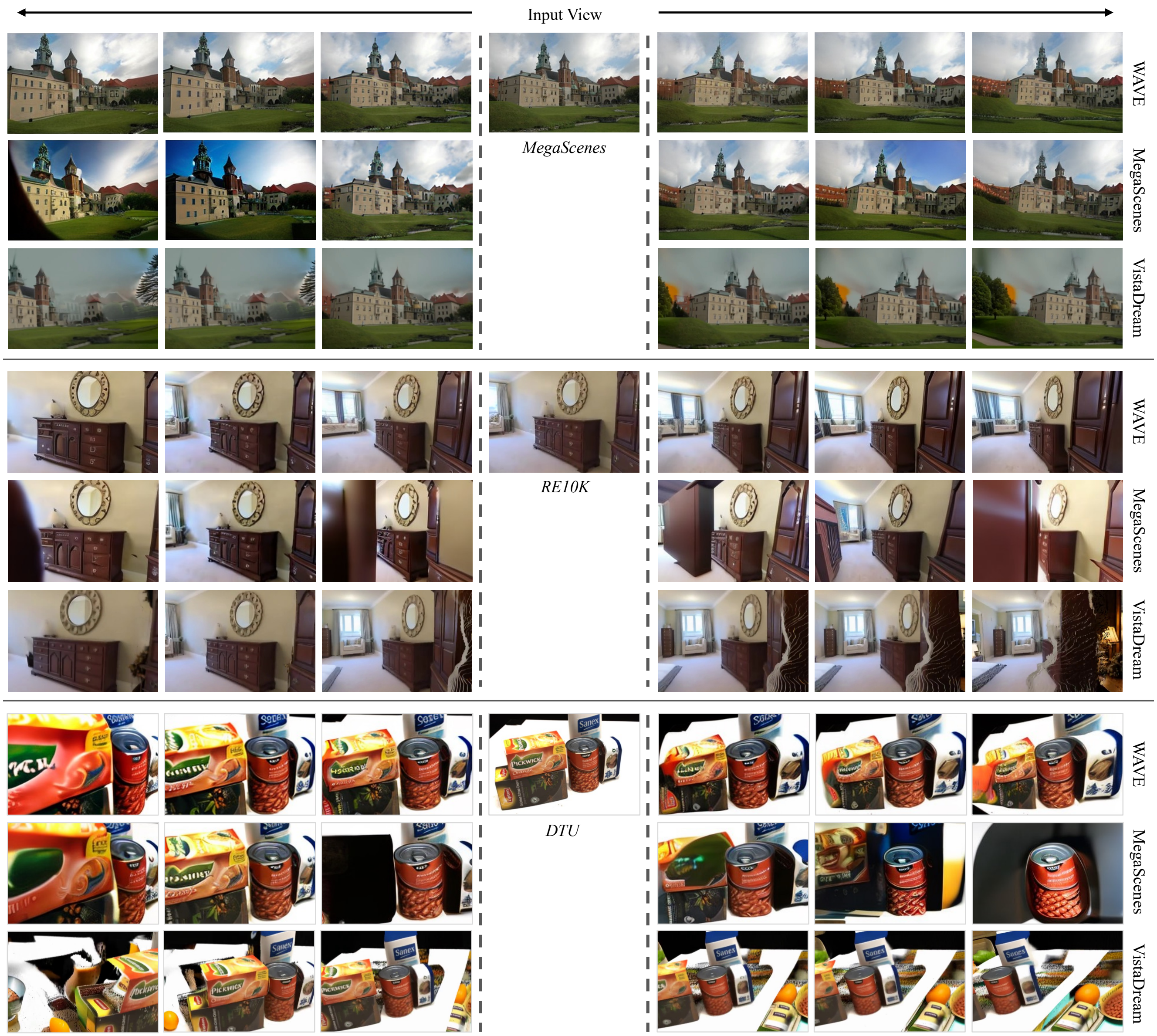

Figure 7. **Qualitative results.** We provide samples for our method and baseline models [25, 32]. The results are obtained by generating images from an input view and continuous camera poses, selecting representative samples from them. Our method exhibits consistent images among generated images. The middle column presents the input view, while the remaining columns display the generated results.

els. More qualitative results and camera parameter visualization can be found in the supplementary materials.

4.3. Ablation Study

We conduct ablative experiments to study the effect of each component in our method. Ablation experiments are performed on the MegaScenes dataset and MegaScenes pretrained model [32]. The quantitative and qualitative results are shown in Table 3 and Fig. 8, respectively.

Effect of warp-guided adaptive attention. Warp-guided adaptive attention improves performances across all metrics compared to the base model in Table 3. As shown in Fig. 8, it helps maintain the spatial structure of most scenes, although color variations persist. This indicates that our attention mechanism preserves structural features by sharing information from other generative processes. Additionally, a detailed comparison with conventional batch self-attention is provided in the supplementary material.

Effect of pose-aware noise initialization. Applying pose-aware noise initialization also improves performance over the diffusion model in Table 3. This shows that pose-aware noise initialization effectively reflects view transformations,

Table 3. **Ablation results.** We analyze different components of our model to assess their impact on the overall performance. PANI refers to the pose-aware noise initialization in Section 3.4, while WGAA denotes the warp-guided adaptive attention in Section 3.2.

Exp. #	PANI	WGAA	*First* LPIPS ↓	*First* CLIPSIM ↑	*Next* LPIPS ↓	*Next* CLIPSIM ↑	Angular Consistency ↓
1			0.609	0.813	0.607	0.861	20.35
2	✓		0.584	0.819	0.560	0.873	19.76
3		✓	0.581	0.822	0.416	0.918	16.49
4	✓	✓	**0.561**	**0.826**	**0.397**	**0.920**	**15.91**

Figure 8. **Visual comparisons in ablation study.** We compare of the generated results when PANI and WGAA are removed from our method, respectively. We generate images along an orbital trajectory, ranging from a specific leftward angle to a specific rightward angle relative to the input image.

helping to reduce noise randomness. As shown in Fig. 8, while objects fail to appear across generated views, color variations do not occur. These results imply that the randomness in noise contributes to color inconsistencies.

4.4. Downstream Task

Additionally, we validate that the consistent images generated using our method effectively enhance the performance of 3D rendering tasks on various datasets [1, 2, 32, 40]. Since 3D rendering tasks rely on multiple images, ensuring view consistency is crucial for providing high-quality multi-view information to 3D models, ultimately improving 3D rendering performance [16, 28]. Thus, this task can be used to evaluate whether the models generate consistent images. Previous studies [25, 26, 30] on diffusion-based 3D rendering with a single image have explored training 3D models using novel view synthesis diffusion model's guidance. However, this approach makes it challenging to assess the diffusion model itself, as it jointly optimizes the 3D model's loss and diffusion guidance loss.

We conduct an evaluation by directly using the generated outputs from the diffusion model as inputs to a 3D model to assess rendering performance. In this experiment, VistaDream is excluded since its pipeline already includes 3D rendering. 3D rendering experiments are conducted using

Table 4. **3D rendering downstream task.** We evaluate 3D rendering performance using 3D Gaussian Splatting [10] with images generated by our method and baselines. In Mip-NeRF 360, ZeroNVS [25] fails to generate reliable images without a rendering technique, making it unsuitable for 3D rendering.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
MegaScenes			
ZeroNVS	23.38	**0.869**	0.098
ZeroNVS + WAVE	**24.23**	0.852	**0.094**
MegaScenes	22.11	0.798	0.128
MegaScenes + WAVE	**24.69**	**0.836**	**0.096**
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**
Mip-NeRF 360			
ZeroNVS	-	-	-
ZeroNVS + WAVE	-	-	-
MegaScenes	25.07	0.829	**0.074**
MegaScenes + WAVE	**26.30**	**0.839**	0.082

3D Gaussian Splatting [10] with 19 image sets generated by our method and other baseline models [25, 32]. As shown in Table 4, our approach outperforms the original methods. These results indicate that our generated consistent images enhance geometric accuracy and visual coherence in 3D reconstruction and rendering tasks. The potential for future application of our method in diffusion-based 3D rendering tasks is also demonstrated. Results for datasets not included in Table 4 are in the supplementary material.

5. Conclusion

We propose a training-free method to improve view consistency in diffusion models for scene-level novel view synthesis from a single image, offering a practical solution for this task. By leveraging inter-view information through attention and noise reinitialization with warped images, our approach overcomes key limitations of diffusion models, including independent generation and noise randomness. Lastly, we address the lack of reliable evaluation metrics by introducing a comprehensive metric process, providing a valuable benchmark for future research.

Limitations and future work. The use of warped images makes the method sensitive to large viewpoint changes, leading to significant missing regions. An autoregressive approach, generating images within a specific view range and iteratively using them for subsequent generations, could address this issue through gradual warping and offer a promising direction for future work.

Acknowledgement. This work was supported in part by IITP RS-2024-00457882 (National AI Research Hub Project), IITP 2020-0-01361, NRF RS-2024-00345806, NRF RS-2023-00219019, G21004687962, and RS-2024-00403860 (Korea Basic Science Institute, National Research Facilities and Equipment Center).

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, 8

[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, 8

[3] Giannis Daras, Yuval Dagan, Alex Dimakis, and Constantinos Daskalakis. Consistent diffusion models: Mitigating sampling drift by learning to be consistent. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[4] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[5] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 1

[6] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, pages 11808–11826. PMLR, 2023. 1

[7] 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

[8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 3, 5

[9] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6700, 2024. 3

[10] 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. 1, 2, 8

[11] Seungryong Kim. Genwarp: Single image to novel views with semantic-preserving generative warping. In *38th Conference on Neural Information Processing Systems Workshop, NeurIPS 2024*. Neural information processing systems foundation, 2024. 2, 3

[12] Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6775–6785, 2024. 2, 5

[13] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *European Conference on Computer Vision*, pages 515–534. Springer, 2022. 1, 2

[14] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021. 2, 3

[15] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2

[16] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9970–9980, 2024. 8

[17] Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. Improving text-to-image consistency via automatic prompt optimization. *arXiv preprint arXiv:2403.17804*, 2024. 3

[18] William R Mark, Leonard McMillan, and Gary Bishop. Post-rendering 3d warping. In *Proceedings of the 1997 symposium on Interactive 3D graphics*, pages 7–ff, 1997. 2, 3

[19] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[20] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10258–10268, 2024. 2, 3

[21] 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*. 2, 3, 5

[22] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 4

[23] 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. 3

[24] Subham Sekhar Sahoo, Aaron Gokaslan, Chris De Sa, and Volodymyr Kuleshov. Diffusion models with learned adaptive noise. *arXiv preprint arXiv:2312.13236*, 2023. 2, 5

[25] 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. 1, 2, 3, 5, 7, 8

[26] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Hyeonsu Kim, Jaehoon Ko, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023. 8

[27] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023. 2

[28] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation, 2024. 8

[29] Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. 2025. 2, 3

[30] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 1, 8

[31] 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. 2, 3, 4

[32] 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. 1, 2, 3, 5, 7, 8

[33] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[34] 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. 2

[35] Haiping Wang, Yuan Liu, Ziwei Liu, Wenping Wang, Zhen Dong, and Bisheng Yang. Vistadream: Sampling multi-view consistent images for single-view scene reconstruction. *arXiv preprint arXiv:2410.16892*, 2024. 1, 2, 3, 5, 6

[36] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7378–7387, 2023. 3

[37] 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. 2, 3, 5

[38] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667, 2024. 1, 2

[39] 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. 3, 4

[40] 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, 8

[41] 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