GenWarp: Single Image to Novel Views with Semantic-Preserving Generative Warping

Junyoung Seo[1,3 *] **Kazumi Fukuda**[1] **Takashi Shibuya**[1] **Takuya Narihira**[1] **Naoki Murata**[1]
Shoukang Hu[1] **Chieh-Hsin Lai**[1] **Seungryong Kim**[3†] **Yuki Mitsufuji**[1,2†]
[1]Sony AI [2]Sony Group Corporation [3]KAIST AI

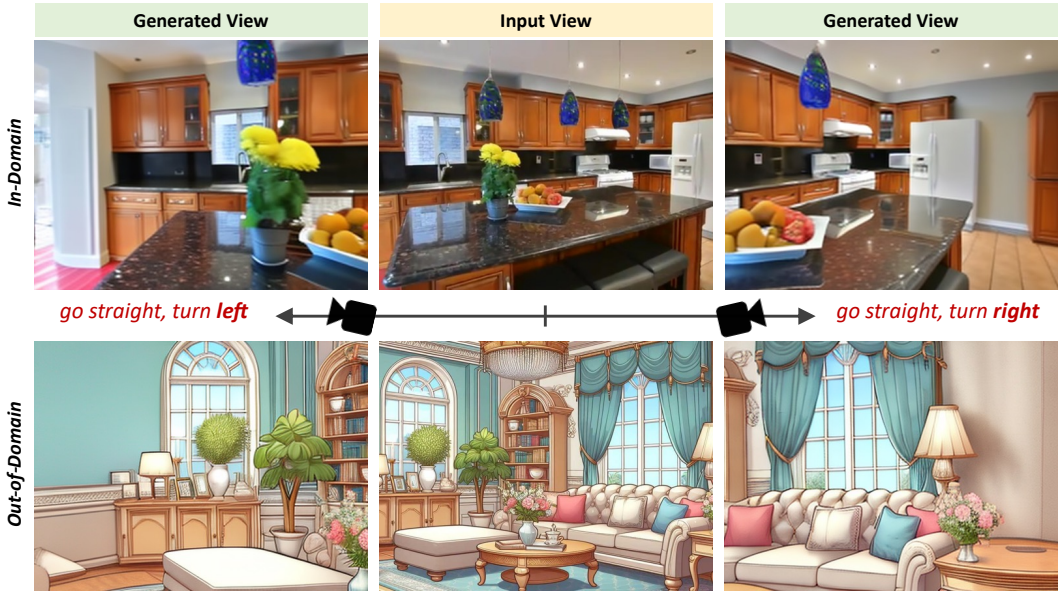

Figure 1: **Teaser.** Our model generates plausible novel views, conditioned on **only a single input view**, enabling to handle both in-domain images (top) and out-of-domain images (bottom).

Abstract

Generating novel views from a single image remains a challenging task due to the complexity of 3D scenes and the limited diversity in the existing multi-view datasets to train a model on. Recent research combining large-scale text-to-image (T2I) models with monocular depth estimation (MDE) has shown promise in handling in-the-wild images. In these methods, an input view is geometrically warped to novel views with estimated depth maps, then the warped image is inpainted by T2I models. However, they struggle with noisy depth maps and loss of semantic details when warping an input view to novel viewpoints. In this paper, we propose a novel approach for single-shot novel view synthesis, a semantic-preserving generative warping framework that enables T2I generative models to learn *where to warp* and *where to generate*, through augmenting cross-view attention with self-attention. Our approach addresses the limitations of existing methods by conditioning the generative model on source view images and incorporating geometric warping signals. Qualitative and quantitative evaluations demonstrate that our model outperforms existing methods in both in-domain and out-of-domain scenarios. Project page is available at `https://GenWarp-NVS.github.io`.

[*]Work done during an internship at Sony AI. [†] Co-corresponding authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1 Introduction

Text-to-image (T2I) diffusion models (*e.g.*, Stable Diffusion [35]) have made rapid progress in generating diverse high-quality images when given a user text prompt. This holds extensive potential utility across various domains, including portrait photo design, cartoon creation, and movie production. However, current T2I models lack the flexibility of moving cameras in the generated image. For example, when a user tries to move the camera closer or farther to the generated image, T2I models often fail to change the viewpoint of the generated image with a proper notion of 3D awareness. This limits its application in real-world scenarios where we hope to achieve user-tailored designing purposes by changing the camera viewpoint for generated images.

To freely move camera viewpoints of an image, a line of research focuses on directly learning a single-shot novel view generation model with a camera viewpoint condition from large-scale 3D datasets. For example, with the advent of large-scale 3D object datasets such as Objaverse [10], recent attempts [26, 40, 27] achieve success in generating novel views of 3D objects from a single image. Beyond the object-centric novel views, efforts for full 3D scenes have also been made [34, 22, 36, 17, 7]. Unlike the object-centric models, these can generate novel views of complex scenes from a single image. The performance of single-shot 3D scene novel view generation models highly depends on the scale of multi-view 3D scene datasets [52, 8, 5]. Compared with the object-centric multi-view datasets [10, 9], it is hard to collect such a large-scale dataset for 3D scenes due to its complexity. Thus, existing models [34, 22, 36, 17, 7] solely trained on these datasets [52, 8, 5] struggle to handle in-the-wild images [17, 36].

Instead of learning dataset-specific novel view synthesis models, alternative approaches [7, 31] propose utilizing the generative prior from large-scale T2I diffusion models, *e.g.*, Stable Diffusion [35]. These works adopt a two-step strategy for novel view generation, called *warping-and-inpainting*, similarly to conventional works [48, 34, 22], with a combination of the large-scale T2I diffusion models and off-the-shelf monocular depth estimation (MDE) models (*e.g.*, MiDaS [33], ZoeDepth [2]). Specifically, they first predict a depth map of a given image via off-the-shelf MDE models [2, 33], and then warp the input image to novel camera viewpoints with the depth-based correspondence, followed by inpainting occluded regions of the warped images with proper text prompts through the T2I diffusion models. The warping-and-inpainting approach successfully generates novel views from in-the-wild images by utilizing large capabilities of T2I diffusion models learned from large-scale image datasets [39].

Despite such an advantage, this warping-and-inpainting approach can generate novel views only in a limited range of camera viewpoints around the input image. This is because **(1) they struggle to handle noisy depth maps predicted by the MDE**. As shown in Fig. 2(a), a reprojection error from the estimated depth map makes the warped image unreliable, becoming a significant performance bottleneck. The subsequent inpainting T2I diffusion models cannot refine the artifacts caused by this error. In addition, **(2) important semantic details of the input view sometimes get lost during geometric warping**, especially when dealing with challenging camera viewpoints. In the above two cases, only a sparse set of pixels is preserved in the warped image, making it difficult to generate the occluded regions while preserving the semantic information of the input view. Fig. 2(b) shows a clear example of this problem; the inpainted regions show a different context with the input view.

To address these issues, we propose a generative warping framework, **GenWarp**, in which we make T2I generative models learn *where to warp* and *where to generate* in images, instead of inpainting unreliable warped images. Our generative model integrates view warping and occlusion inpainting into a unified process, unlike existing two-step approaches that perform these operations separately. By directly taking the input view images with their estimated depth maps, our model learns to warp them and to generate the occluded or ill-warped parts with our augmented self-attention. Our approach eliminates the dependency on unreliable warped images and integrates semantic features from the source view, preserving the semantic details of the source view during generation. Similar to [26, 40, 49], we leverage generalization capabilities of T2I diffusion models (*e.g.*, Stable Diffusion [35]) through fine-tuning a T2I diffusion model.

Our main contributions are as follows:

- We propose a semantic-preserving generative warping framework, **GenWarp**, to generate high-quality novel views from a single image.

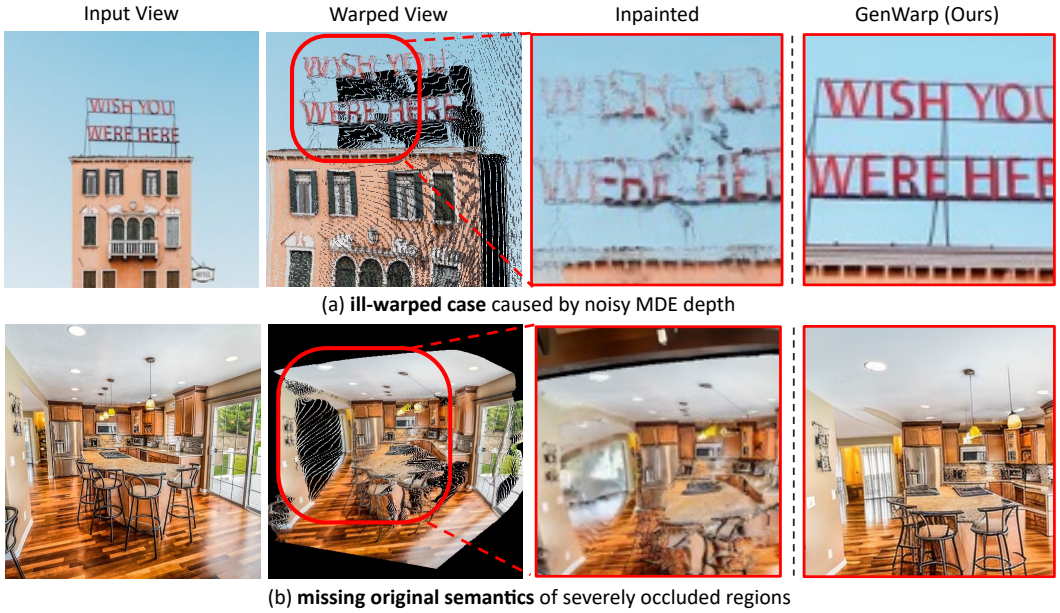

Figure 2: **Limitations of explicit warping-and-inpainting approach [35, 7, 31].** Results from challenging new camera viewpoints for warping-and-inpainting approach show artifacts. (a) The neon sign present in the input view is distorted after geometric warping due to the noisy depth. (b) The next room peeked in from the new camera viewpoint lacks the context given by the input view.

- **GenWarp** learns *where to warp* and *where to generate* in images through augmenting self-attention with cross-view attention instead of inpainting unreliable warped images, which eliminates the artifacts caused by error depths and integrates semantic features from source views, preserving semantic details in generation.
- Extensive experiments on RealEstate10K [52], ScanNet [8], and in-the-wild images (*e.g.*, AI-generated images) validate that **GenWarp** achieves superior performances over existing methods in both in-domain and out-of-domain scenarios.

2 Related Work

Generating novel views from a single image is a challenging ill-posed problem that has primarily been addressed in combination with generative modeling. These novel view generative models can generally be categorized into two types: those designed for object-centric scenes and those designed for general scenes, including indoor and outdoor scenes. On the other hand, following the recent success of large-scale T2I models, there are methods that can control the generation results, exploiting the attention mechanism.

Single-shot novel view synthesis for objects. Following the success of image diffusion models [15, 35], diffusion models for novel view synthesis [47, 4] have been proposed. These works train diffusion models to take a single image and a novel camera viewpoint as conditions, and directly generate novel view images. More recently, with the emergence of large-scale 3D datasets such as Objaverse [10, 9], generalized generative models for single-shot novel view synthesis (NVS) have emerged. Recent works [26, 27, 19], including Zero123 [26], achieves powerful generalization capability by fine-tuning T2I diffusion models on Objaverse. While these models enable object-centric novel view synthesis from an in-the-wild single image, such generalized novel view models for general scenes remain relatively unexplored.

Single-shot novel view synthesis for general scene. Single-shot NVS often necessitates generating outer regions or occluded regions that are not visible in an input view. Thus, recent works [48, 34, 25, 22] propose generating novel views in a warping-and-refining fashion, which involves first predicting a depth map of an input view, then warping the input view along with the depth map to a desired viewpoint, and finally refining missing regions arising from the geometric warping. Another line of works [36, 44, 17] directly train novel view generative models without depth-based warping. For

example, GeoGPT [36] achieves novel view generation, by feed-forwarding an input view and a camera viewpoint to a transformer-based architecture. Other recent works [44, 17] train novel view diffusion models with a cross-view attentions [44, 17], or an epipolar constraint [44]. More recently, some works [31, 7, 41] have proposed bringing a large-scale T2I model [35] to the warping-and-refining strategy. This approach enables the generation of novel views from in-the-wild images, which was previously challenging. Nonetheless, it shows unstable results, especially when the camera viewpoint is far from its original position. Concurrently, ZeroNVS [38] fine-tunes Zero123 [26] for NVS in general scenes. It focuses on camera parametrization to avoid 3D scale ambiguity, which differs from our focus; we focus on improving depth warping-based NVS.

Attention-based control in large-scale T2I models. Since the emergence of large-scale text-to-image (T2I) diffusion models [35, 37], recent works [21, 40, 3, 14, 16] have investigated the properties of self-attention within T2I models. For example, Text2Video-Zero [21] and MVDream [40] generate consistent images by sharing self-attention between video frames or 3D multi-views, respectively. Similarly, Animate-Anyone [16] and MagicAnimate [49] generate human dance videos through a fine-tuned T2I model that shares self-attention with input image features. Observing the generalization capability and efficiency of these self-attention-based controllable architectures, our approach is highly influenced by them. However, using these architectures for single-shot novel view generation is non-trivial, as input scenes are usually complex, and the details within them must be generated consistently with the camera movements. Our model integrates MDE depth-based correspondence while benefiting from the advantages of these architectures, thereby significantly improving single-shot NVS performance.

3 Method

3.1 Preliminaries and problem statement

Given a single image for an input view I_i, our goal is to generate a novel view I_j from a relative camera viewpoint $P_{i\to j}$ and a camera intrinsic K. To enable this, recent works [7, 31] propose a *warping-and-inpainting* framework that adopts monocular depth estimation (MDE) models [2] for geometric warping and T2I generative models [35, 37] for inpainting. In this approach, an input view image I_i is geometrically warped with its MDE depth map D_i to the desired camera viewpoint $P_{i\to j}$:

$$I_{\text{warp}} = \text{warp}\big(I_i; D_i, P_{i\to j}, K\big), \tag{1}$$

where $\text{warp}(\cdot)$ is a geometric warping function which unprojects pixels of an input view image I_i with its depth map D_i to 3D space, and reprojects them based on the desired camera conditions, $P_{i\to j}$ and K. More specifically, in homogeneous coordinates, pixel location x_i in the input view is transformed to the pixel location x_j in the novel view such that:

$$x_j \simeq K P_{i\to j} D_i(x_i) K^{-1} x_i, \tag{2}$$

where the projected coordinate x_j is a continuous value. To obtain the warped image I_{warp}, it is followed by mapping the pixel colors from the input view to the novel view with a flow field between x_i and x_j [48, 30, 34]. And, inpainting diffusion model ϕ generates a novel view image by filling the missing regions in the warped image I_{warp}:

$$I_j \sim p_\phi(I_j; I_{\text{warp}}, M_{\text{warp}}, c_j), \tag{3}$$

where $p_\phi(\cdot)$ is a learned distribution of the diffusion model ϕ conditioned on an occlusion mask M_{warp}, a text prompt c_j, and the warped image I_{warp} used for inpainting. This approach assumes that the estimated depth map D_i is accurate. Under the assumption, the warped image I_{warp} would be an ideal guidance for generation.

However, we have observed that the warped image is often not reliable when a novel camera viewpoint is far from the original viewpoint. It is because this explicit warping operation is sensitive to errors in the depth map, and depth maps predicted by MDE are usually noisy, arising artifacts after the warping. The subsequent inpainting model only takes the warped image I_{warp} as input which contains ill-warped artifacts outside the region to be inpainted, thus showing limited performance at large view changes. Additionally, the warped image may lose the semantic information originally contained in the input view due to factors such as occlusion, but this approach does not take that into account, as exemplified in Fig. 2.

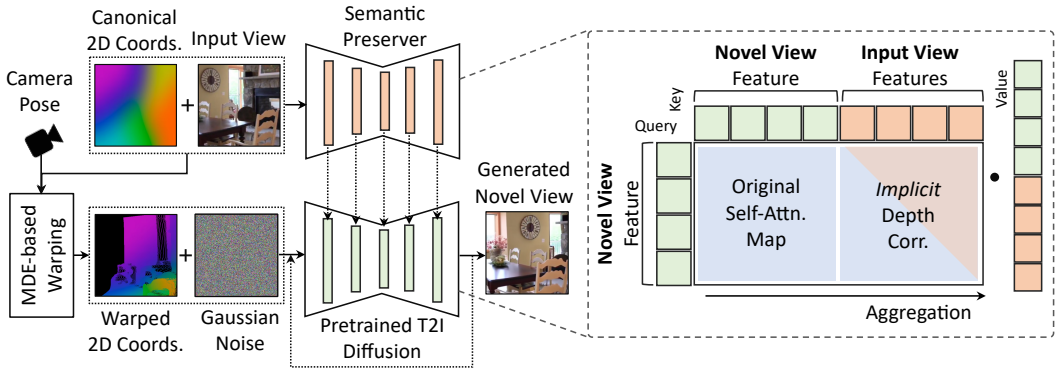

Figure 3: **Method overview:** (**Left**) Given an input view and a desired camera viewpoint, we obtain a pair of embeddings: a 2D coordinate embedding for the input view, and a *warped* coordinate embedding for the novel view from estimated depth through MDE. With these embeddings, a semantic preserver network produces a semantic feature of the input view, and a diffusion model conditioned on them learns to conduct geometric warping to generate novel views. (**Right**) We augment self-attention with cross-view attention, followed by aggregating the features with both attentions at once. It helps the model to consider where to generate and where to warp.

3.2 Semantic-preserving generative warping

To alleviate the aforementioned limitations, we introduce a novel approach where a diffusion model learns to implicitly conduct geometric warping operation, instead of warping the pixels or the features directly. We design the model to interactively compensate for the ill-warped regions during its generation process, thereby preventing artifacts typically caused by explicit warping. In addition, to preserve the semantics in the input view, our framework takes the input view image without warping, and the encoded semantic features of the input view are incorporated into the generation process, which is different from other approaches that solely take an unreliable warped image from which the original semantics are difficult to infer.

To this end, we leverage the attention layers inside the pre-trained diffusion U-net [35]. Our key idea is to learn the attention between input view and novel view features, which serves as an *implicit* correspondence that mimics explicit depth-based warping within the diffusion model. By incorporating this into the diffusion process, we aim to seamlessly integrate the effect of depth-based warping into the generative prior. This implicit correspondence in the form of attention can be integrated into the existing self-attention layers inside the diffusion U-net. In so doing, the input view features additionally interact with the novel view features in the generation process, making the diffusion models naturally find *where to generate* and *where to warp*, as visualized in Fig. 4.

Two-stream architecture. Our approach comprises a two-stream architecture, a semantic preserver network and a diffusion model, sharing an identical U-net-based architecture. The semantic preserver network takes the input view image I_i and produces a semantic feature F_i of the input view. And, the diffusion model generates a novel view image I_j, by integrating the input view feature F_i into its internal novel view feature F_j. To imbue the diffusion model with the MDE depth-based correspondence, we use a pair of canonical coordinates and warped coordinates as additional conditions. Fig. 3 illustrates an overview of our architecture. In the following, we explain each component in detail.

Warped coordinate embedding. To condition on the MDE depth-based correspondence, we use two coordinate embeddings, a canonical coordinate embedding for the input view, and a *warped* coordinate embedding for the novel view. We are motivated to use the warped coordinate embedding by [29], whose purpose is correspondence-based appearance manipulation. Here, we extend this concept to the geometric warping for novel view generation.

Specifically, we construct a canonical 2D coordinate map $X \in \mathbb{R}^{h \times w \times 2}$, where each value is normalized between -1 and 1. This 2D coordinate map is transformed by a positional encoding function γ into Fourier features [43] $C_i = \gamma(X)$. We use this Fourier feature map C_i as the coordinate embedding for the input view I_i. We then geometrically warp this coordinate embedding C_i of the

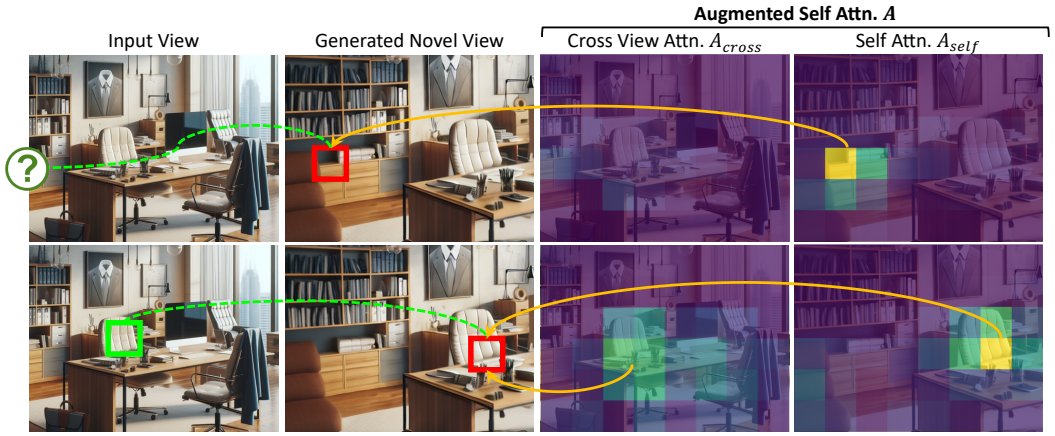

Figure 4: **Visualization of augmented self-attention map.** In augmented self-attention map A, the original self-attention part A_{self} is more attentive to regions requiring generative priors, such as occluded or ill-warped areas **(top)**, while the concatenated cross-view attention part A_{cross} focuses on regions that can be reliably warped from the input view **(bottom)**. By aggregating both attentions at once, the model naturally determines which regions to generate and which to warp.

input view to the desired novel viewpoint $P_{i\to j}$:

$$C_j = \text{warp}(C_i; D_i, P_{i\to j}, K), \tag{4}$$

where $\text{warp}(\cdot)$ is the same geometric warping function in Eq. 1. The warped coordinate embedding C_{j} serves as the coordinate embedding for the novel view I_j. These coordinate embedding C_i for the input view and C_j for the novel view are added to the source view feature F_i and the target view feature F_j through convolution layers respectively. This embedding strategy guides the model to follow the geometric correlation between the input view and the novel view. In an explicit warping strategy, the process is inherently affected by depth estimation errors in virtue of regarding the warped image as a reliably given condition. However, by implicitly learning to warp with this embedding, it is expected that the influence of these errors can be mitigated.

Augmenting self-attention with cross-view attention. To infuse the input view features F_i, we first construct a cross-view attention, where the cross-view attention map represents the similarities between the input view and the novel view being generated. Thanks to the coordinate embeddings, the cross-view attention map learns to give depth-based correspondence that can be absorbed into the generative process. Then, we propose concatenating this cross-view attention to the existing self-attention.

Specifically, we concatenate the keys and values of self-attention layers in the diffusion U-net with the input view features F_i, and apply the self-attention [45] with the following query, key, and value:

$$q = F_j, \quad k = [F_i, F_j], \quad v = [F_i, F_j], \tag{5}$$

where F_i is the input view feature in the semantic preserver network, and F_j is the novel view feature in the diffusion U-net. Then we can obtain the augmented self-attention map A, which is a concatenation of the cross-view attention map A_{cross} and the self-attention map A_{self}.

By aggregating the values with both attentions at once, the model learns to balance the contributions from the novel view's self-attention A_{self} and the cross-view attention A_{cross}. Our intuition behind this design is that the original self-attention A_{self} in the diffusion U-net attends to the generative prior, and the cross-view attention A_{cross} attends to the warping prior from the input view. This allows the model to inherently decide which regions should rely more on its generative capability and which areas should depend primarily on the information from the input view warping, as shown in Fig. 4.

3.3 Training strategy

Fine-tuning pretrained text-to-image diffusion models. We leverage the pretrained Stable Diffusion 1.5 model [35] for both diffusion U-net and semantic preserver network, to inherit its generalization capability. Different from Stable Diffusion which takes text prompt embedding through

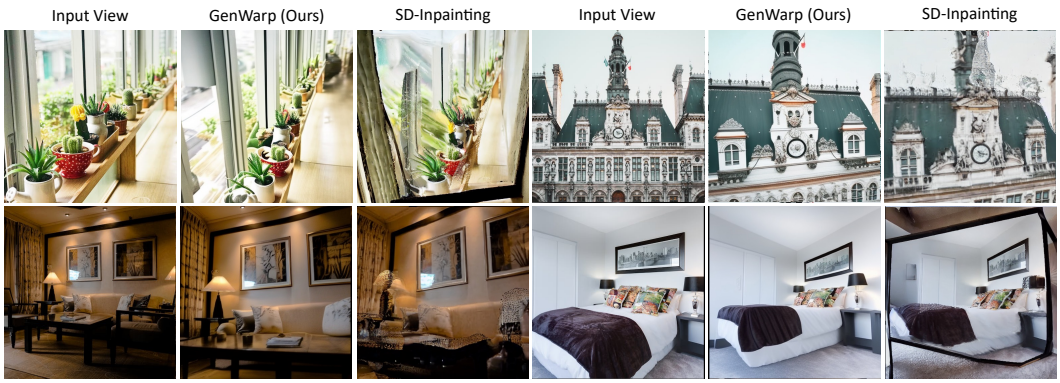

Figure 5: **Qualitative results with images in the wild.** We compare our method with Stable Diffusion Inpainting [35] on in-the-wild images. More qualitatvie results can be found in Fig. 10 of Appendix.

Figure 6: **Consistent view generation results.** Our model can generate consistent multiple views by taking pre-generated novel views as inputs.

CLIP [32], our model takes an image and a desired camera viewpoint as inputs. Therefore, we replace text condition needed for Stable Diffusion with image embedding of the input image through a CLIP image encoder. As all the components in our framework can be trained in an end-to-end manner, we use a sole training loss for fine-tuning, which is the same as the original training loss in LDM [35]. Given a dataset $\mathcal{X}$ consisting of pairs of source view image I_i, target view image I_j, their camera information $P_{i\to j}$, and a depth map D_i, we first encode the source view I_i and the target view image I_j to their corresponding latents z_i and z_j through the LDM encoder, respectively. Then the model is fine-tuned using the following loss function:

$$\mathcal{L}_{\text{ours}}(\theta,\psi) = \mathbb{E}_{\mathcal{X},t,\epsilon}\left[\|\epsilon - \epsilon_{\theta,\psi}(z_{j,t}; z_i, D_i, P_{i\to j}, K)\|_2^2\right], \tag{6}$$

where $z_{j,t}$ denotes a noised latent of z_j at diffusion timestep t. $\epsilon_{\theta,\psi}(\cdot)$ is our model including the diffusion U-net θ and the semantic preserver network ψ, which predicts the added noise in the diffusion process.

Data preparation. We fine-tune the model on multi-view datasets including indoor scene and outdoor scene, *i.e.*, RealEstate10K [52], ScanNet [8], ACID [25]. Specifically, we sample two consecutive frames at intervals of 30-120 frames to make pairs of source view and target view images. For ScanNet [8], we use provided ground-truth depth maps and the camera information. For RealEstate10K [52] and ACID [25], ground-truth depth maps are not provided. So, we pre-process the datasets to generate pseudo ground-truth depth maps and their corresponding camera information. Specifically, we use DUSt3R [46] as a pair-depth estimator, followed by PnP-RANSAC [13, 23] to find the corresponding camera information aligned with the estimated depth maps. Additionally, we exclude pairs with low-confident depth maps in our training dataset. Note that our model is less affected by 3D scale ambiguity [6, 38] in this procedure, as camera parameters are aligned to the scales of estimated depth maps.

4 Experiments

4.1 Experimental setup

We train our model on multiple datasets, including indoor RealEstate10K [52], ScanNet [8], and outdoor ACID [25] datasets. For fair quantitative and qualitative comparison with baseline methods [36, 17] trained on RealEstate10K [52], we also prepared a version of our model trained on the same single dataset. Our baseline methods include GeoGPT [36], Photometric-NVS [17], and the

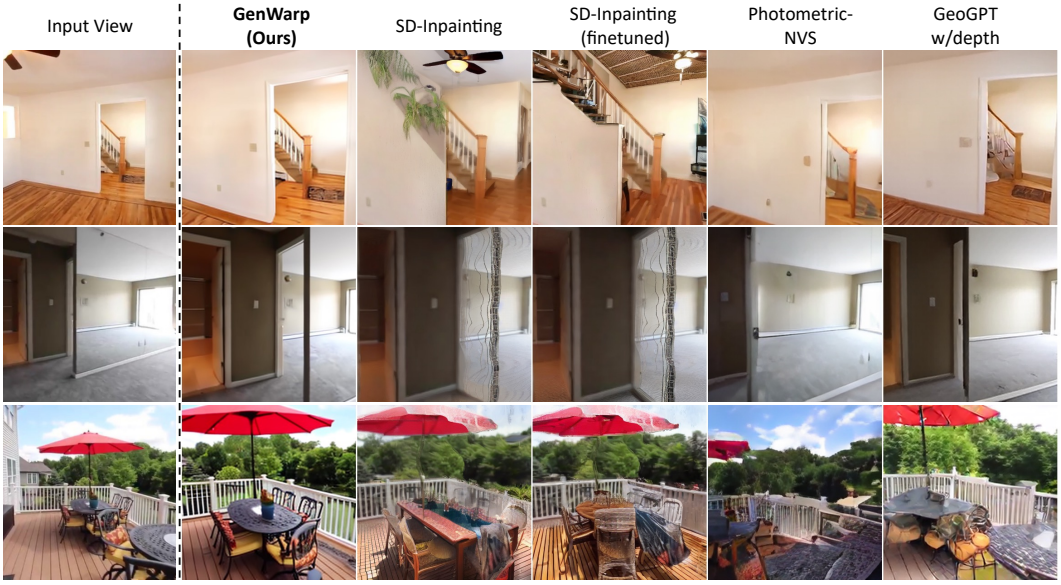

Figure 7: **Qualitative comparisons with baseline methods [35, 17, 36].** We present single-shot novel view generation results with large viewpoint changes on RealEstate10K [52] test set. Our GenWarp generates high-quality novel views consistent with the input views. We also provide qualitative results on ScanNet [8] in Fig. 9 of Appendix.

warping-and-inpainting method [7, 31] using the Stable Diffusion Inpainting model [35]. To ensure a fair evaluation, we also provide results of the Stable Diffusion Inpainting model fine-tuned on the same multi-view dataset [52]. For GeoGPT, we compare with the results from the depth conditional setting, which is most similar to our approach. For the methods that utilize depth information, namely, our method, GeoGPT, and Stable Diffusion Inpainting, we use the same monocular depth estimation models [2, 46]. Please refer to Appendix B for additional details.

4.2 Qualitative results

Qualitative results on in-the-wild images. Fig. 5 and Fig. 6 show qualitative results on in-the-wild images, *e.g.*, cartoonish pictures, real photos, and AI-generated [1] images. SD-Inpainting [35]-based warping-and-inpainting approach, used in recent works [7, 31], shows reasonable results with a good generalization capability for extrapolation, but ill-warped artifacts exist in some areas. In contrast, our method consistently generates feasible novel views by refining those artifacts well.

Qualitative comparisons. We present qualitative comparisons with the baseline methods [35, 17, 36] in Fig. 7. The warping-and-inpainting approaches with the SD-Inpainting model [35, 7, 31] show good performance for areas where the input view and novel view clearly overlap. However, for regions where warped pixels are sparse, it generates inconsistent novel views without considering the semantic information of the input view. Photometric-NVS [17] and GeoGPT [36] show reasonable performance when the camera view changes are small. However, their performance degrades when the view change is large or when the given images of scenes are underrepresented in the training data, such as outdoor scenes. Our method generates plausible novel views and is robust to variations in the type of scenes and camera viewpoints, by considering the semantics of the input views.

4.3 Quantitative results

We perform a quantitative comparison of our model and baseline models [35, 36, 37] trained on RealEstate10K [52], on the test set of RealEstate10K (in-domain) and ScanNet [8] (out-of-domain) using FID for generation quality on distribution level and PSNR for reconstruction quality, with 1,000 generated images. We categorize the distance between source and target views into mid range (30-60 frames) and long range (60-120 frames). Tab. 1 demonstrates that our method shows superior performance in both out-of-domain setting and in-domain setting. The SD-Inpainting-based approaches perform well in terms of PSNR thanks to explicit warping, but struggle with ill-warped artifacts resulting in poor FID. GeoGPT shows good generation quality as evidenced by its FID score

Methods	Out-of-domain [8]		In-domain [52]			
	Mid range		Mid range		Long range	
	FID ↓	PSNR ↑	FID ↓	PSNR ↑	FID ↓	PSNR ↑
GeoGPT [36] w/depth	85.52	11.36	<u>32.70</u>	12.26	<u>33.91</u>	11.69
Photometric-NVS [17]	N/A	N/A	37.17	12.05	39.93	11.63
SD-Inpainting [35, 7]	<u>52.20</u>	<u>11.68</u>	41.76	14.21	44.13	12.98
SD-Inpainting [35, 7] (fine-tuned)†	72.90	9.10	39.17	<u>14.35</u>	43.08	<u>13.10</u>
GenWarp (Ours)	**46.03**	**12.95**	**31.10**	**14.55**	**32.40**	**13.55**

Table 1: **Quantitative comparisons.** We compare our method with novel view generative models [36, 17] and warping-and-inpainting approach consisting of Stable Diffusion Inpainting [35, 7, 31], on in-domain setting (training dataset [52]), and out-of-domain setting (external dataset [8]). † We additionally provide results of Stable Diffusion Inpainting fine-tuned on the multi-view dataset [52].

but tends to disregard input view details, leading to poor PSNR. In the out-of-domain setting, the SD-Inpainting shows reasonable performance, but its performance deteriorates after fine-tuning on the multi-view dataset [52]. For Photometric-NVS [17], we exclude its out-of-domain results as its provided model trained on RealEstate10K [52] fails to generate novel views when camera parameters in ScanNet [8] are given.

4.4 Ablation study

Embeddings for warping signal. We perform an ablation study on the embeddings used to create the warping signal by geometric warping with MDE depth maps. We compare the warped coordinate embedding to other possible candidates: warped depth and warped image. Additionally, we test the camera embedding as condition. Specifically, we encode the camera viewpoint to a Plücker ray representation [42] and replace the warped coordinate embedding with this camera embedding. As shown in Tab. 2, the warping signal given by the warped coordinate embedding is the most effective among them.

Conditions	FID ↓
Warped coordinates	**32.40**
Warped depth map	34.17
Warped image	35.27
Camera embedding [42]	39.10

Table 2: **Ablation on embeddings.**

Camera viewpoint variations. Fig. 8 illustrates the relationship between the difficulty of camera viewpoint changes and the degree of distortion in the generated novel views. We adopt the analysis of view changes proposed in [36], using the LPIPS [51] metric between GT source and target views as a proxy for viewpoint change difficulty, and the LPIPS between the generated and GT target views as a measure of distortion. As shown in Fig. 8, our method achieves the least distortion compared to the baseline methods. Inpainting-based methods [35] show the second-best performance when the viewpoint change is not large, but GeoGPT [36] shows better performance in the case of extreme viewpoint changes.

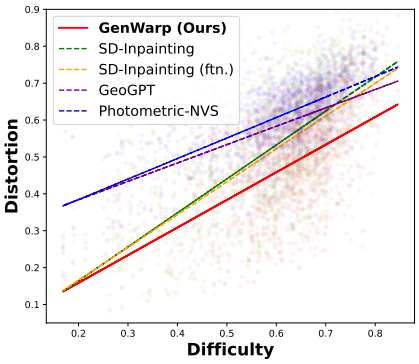

Figure 8: **Comparison on various viewpoint changes.**

5 Conclusion

We have proposed **GenWarp**, a framework for generation of novel views from a single image, preserving semantics contained in the input view by learning to warp images through a generative process. By augmenting the self-attention in diffusion models with cross-view attention conditioned on the warping signal, our approach learns to preserve the semantics of the input view while naturally determining where to warp and where to generate. Extensive experiments demonstrate that GenWarp generates higher-quality novel views compared to existing methods, especially for challenging viewpoint changes, while exhibiting generalization capability to out-of-domain images.

Societal Impacts

This paper presents in the field of AIGC (AI-Generated Content). The proposed model in the paper generates images of user-provided camera viewpoints based on input images. Therefore, while there may be potential social impacts as a consequence, there is nothing in particular to be highlighted. Our model relies on learning from large-scale multi-view datasets, so it may reflect potential societal biases included in these datasets.

Acknowledgment

This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, RS-2024-00509279) and the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025, RS-2023-00266509, RS-2024-00333068), and National Research Foundation of Korea (RS-2024-00346597).

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023.

[4] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4217–4229, 2023.

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[6] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. *arXiv preprint arXiv:2312.12337*, 2023.

[7] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023.

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[9] 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.

[10] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023.

[11] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20637–20647, 2023.

[12] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2024.

[13] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[14] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. *arXiv preprint arXiv:2312.02133*, 2023.

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

[16] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023.

[17] J Yu Jason, Fereshteh Forghani, Konstantinos G Derpanis, and Marcus A Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7071–7081. IEEE, 2023.

[18] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014.

[19] Yash Kant, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, Igor Gilitschenski, and Aliaksandr Siarohin. Spad: Spatially aware multiview diffusers. *arXiv preprint arXiv:2402.05235*, 2024.

[20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

[21] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023.

[22] Jing Yu Koh, Harsh Agrawal, Dhruv Batra, Richard Tucker, Austin Waters, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Simple and effective synthesis of indoor 3d scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1169–1178, 2023.

[23] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, 81:155–166, 2009.

[24] 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.

[25] 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.

[26] 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.

[27] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023.

[28] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[29] Jiteng Mu, Shalini De Mello, Zhiding Yu, Nuno Vasconcelos, Xiaolong Wang, Jan Kautz, and Sifei Liu. Coordgan: Self-supervised dense correspondences emerge from gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10011–10020, 2022.

[30] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5437–5446, 2020.

[31] Hao Ouyang, Kathryn Heal, Stephen Lombardi, and Tiancheng Sun. Text2immersion: Generative immersive scene with 3d gaussians. *arXiv preprint arXiv:2312.09242*, 2023.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[33] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

[34] Chris Rockwell, David F Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14104–14113, 2021.

[35] 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.

[36] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14356–14366, 2021.

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[38] 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 real image. *arXiv preprint arXiv:2310.17994*, 2023.

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[40] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

[41] Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*, 2024.

[42] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34:19313–19325, 2021.

[43] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

[44] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16773–16783, 2023.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[46] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. *arXiv preprint arXiv:2312.14132*, 2023.

[47] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022.

[48] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7467–7477, 2020.

[49] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*, 2023.

[50] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[52] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

Appendix

A Additional qualitative results

Fig. 9 shows qualitative results on out-of-domain setting, *i.e.*, testing on ScanNet [8] with our method and baseline methods [36, 35, 7] trained on RealEstate10K [52]. We also provide additional qualitative results on in-the-wild images in Fig. 10.

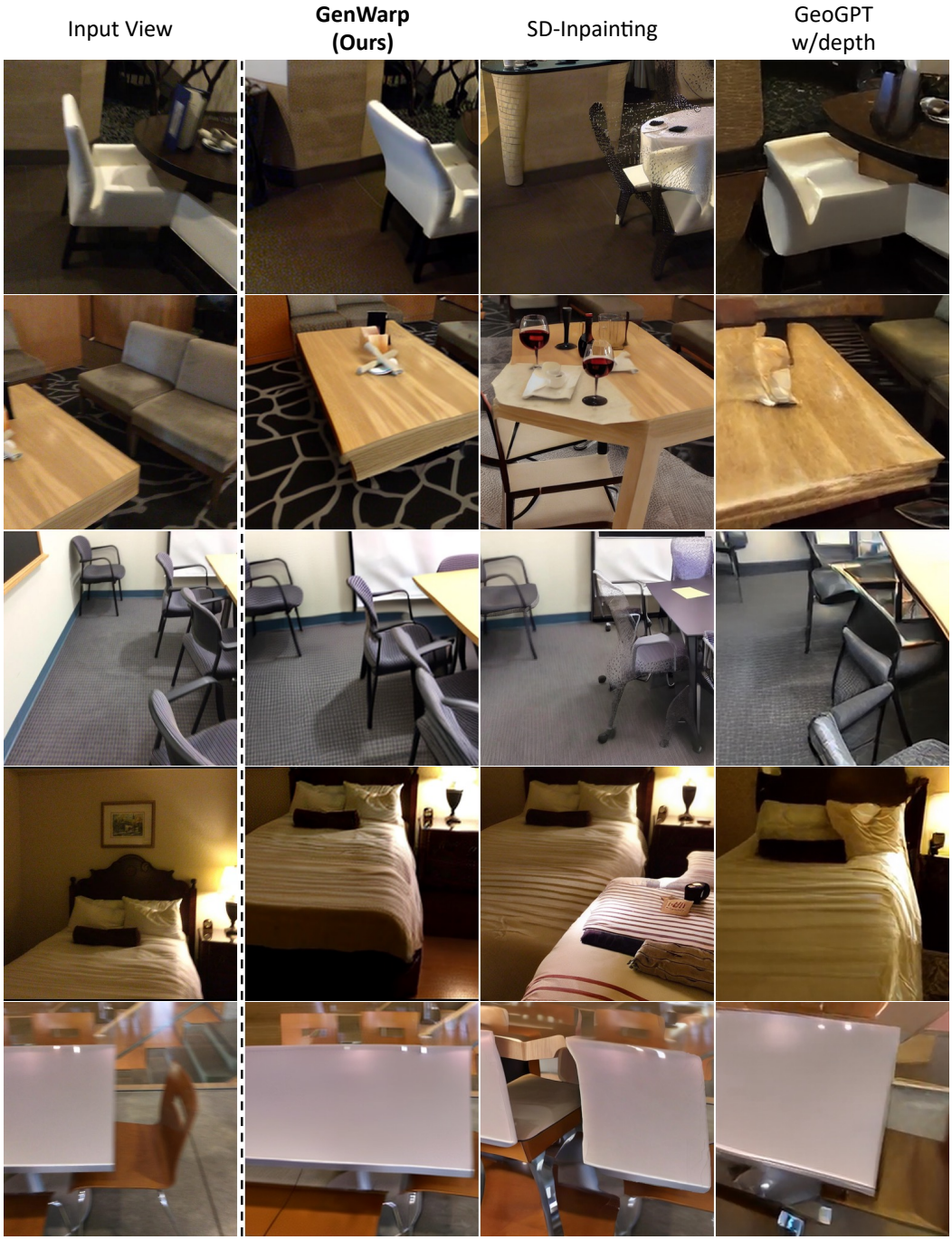

Figure 9: **Extensive qualitative comparisons in out-of-domain setting.** We provide qualitative results of our model trained on RealEstate10K [52], on the external dataset, ScanNet [8].

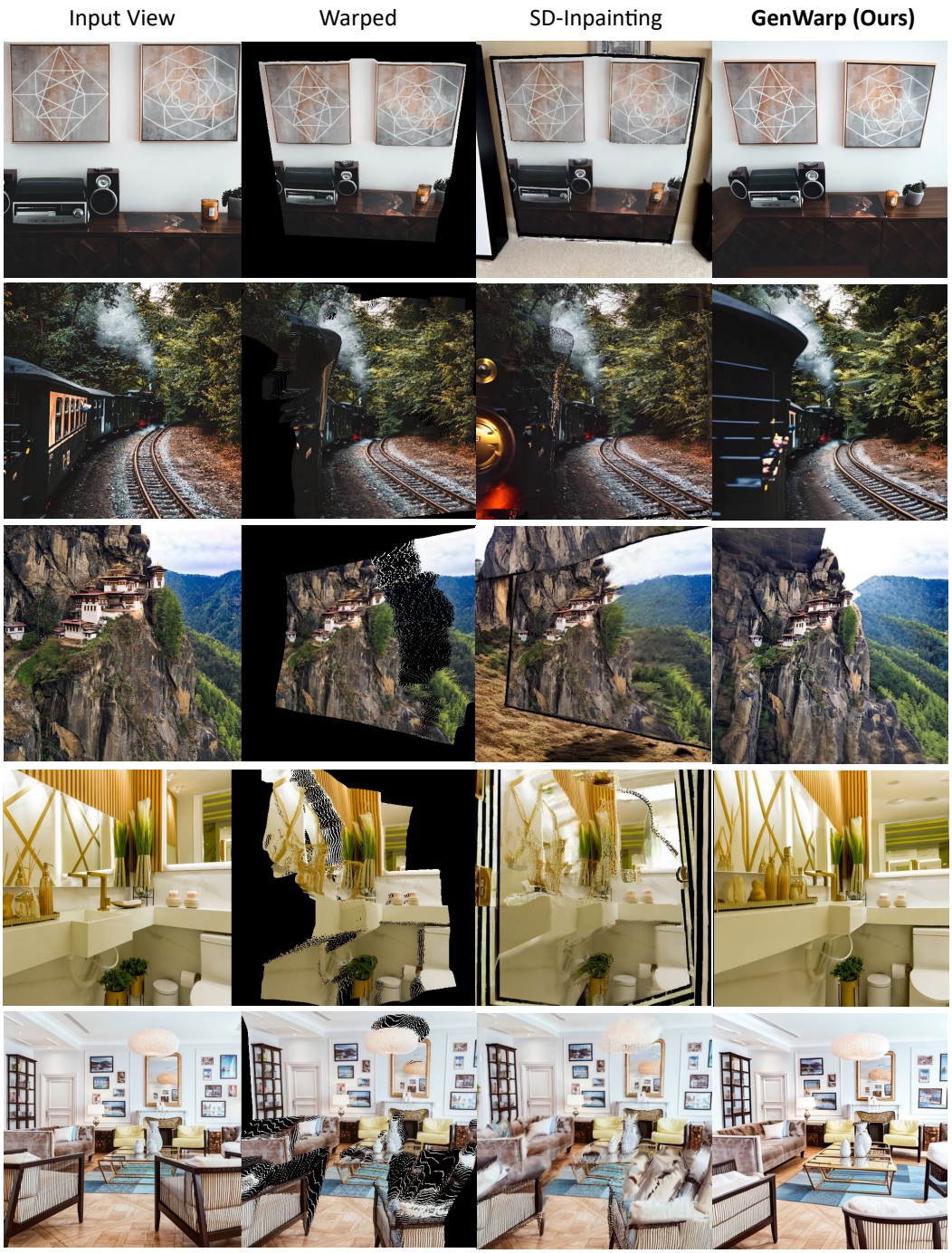

Figure 10: **Extensive qualitative results on in-the-wild images.** We present extensive qualitative results of our method and baseline methods [35, 7] on the in-the-wild images.

B Additional implementation details

We initialize our two networks, semantic preserver and diffusion U-net, with Stable Diffusion v1.5 [35], and fine-tune the networks on 2×H100 80GB with a batch size of 48 for 2-3 days, at resolutions of 512×384 and 512×512. Specifically, we fine-tune the whole parameters of the semantic preserver network and the diffusion U-net in an end-to-end manner. All the hyper-parameters used in our training are kept same as the training of Stable Diffusion 1.5 [2]. In inference, it takes around 2 seconds to generate a novel view with a single H100 80GB.

[2]Stable Diffusion v1.5 Model card: https://huggingface.co/runwayml/stable-diffusion-v1-5

Monocular depth estimation. We use two external depth estimation networks for all qualitative and quantitative results: ZoeDepth [2] and DUSt3R [46]. DUSt3R is a model that predicts pointmaps given two images as a pair. We use the z-values of these pointmaps in two ways: as a pair depth estimation during training and as a monocular depth estimation during inference (by using the same image for both input views). For quantitative evaluation, we use DUSt3R for depth prediction as the predicted depth maps are passed through the same normalization in the process of DUSt3R with the pseudo depth pairs in training dataset which are estimated using the same network. For qualitative comparisons, we use ZoeDepth to predict metric depth maps. Note that we have used the same estimated depth maps for our method and all the baseline methods [36, 35, 7] which need depth information.

Reproducing warping-and-inpainting approach with T2I inpainting models. To implement the warping-and-inpainting strategy using Stable Diffusion Inpainting [36], we follow 'Dream' stage of LucidDreamer [7], which consists of inpainting using the pretrained T2I model [36] after depth-based warping via monocular depth estimation in the official code repository. We observe that directly applying the occlusion mask from depth-based warping to the Stable Diffusion Inpainting model leads to the generation of collapsed images. As suggested in the official code of LucidDreamer, increasing the occlusion mask size for occlusions below a certain threshold effectively prevents this collapse. However, this approach involves a trade-off, as it may further ignore pixels from the source view. Additionally, when using challenging camera trajectories (especially when moving the camera forward), artifacts still occur despite this mask filtering. To address this, we set the minimum occlusion size to 8×8 and expand smaller occlusions to this size, considering that a resolution of latents in LDM [35] is 8 times lower than that of the images. We use inverse warping for the existing warping-and-inpainting method, which provides natural interpolation and reduces occlusion. In contrast, our method employs forward warping to facilitate the intervention of the generative prior. Fig. 11 shows the difference between forward warping and inverse warping, and the obtained occlusion masks which are used in the subsequent inpainting procedure.

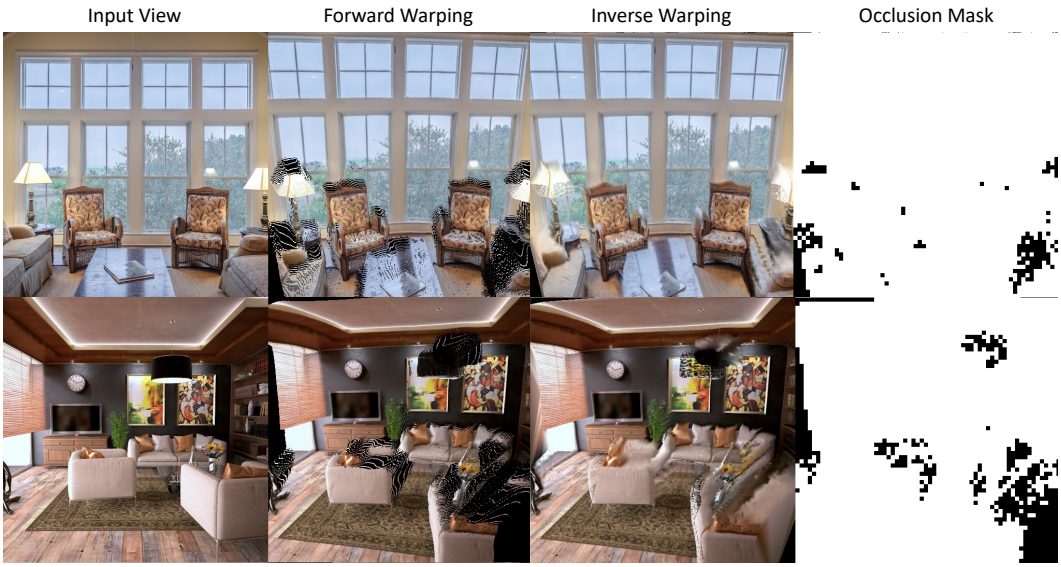

Figure 11: Following LucidDreamer [7], we apply inverse warping and occlusion mask filtering to reproduce the existing warping-and-inpainting approach [7, 31] with Stable Diffusion Inpainting [35].

C Additional discussion

Explicit feature warping vs. implicit warping (ours). Another straightforward approach for integrating depth-based warping into diffusion models is to warp features within the diffusion model's feature space. We initially tried this diffusion feature warping. Specifically, the input view feature F_i from the semantic preserver network is geometrically warped with the corresponding depth map, and then added to the diffusion U-net's intermediate feature F_j through zero-initialized convolution layers.

This approach shows reasonable performance on multi-view datasets like ScanNet [8], which include ground-truth sensor depth. However, most multi-view datasets [25, 52] are derived from videos and lack dense GT depth. To address this, we use estimated depth maps from DUSt3R [46], as described in Sec. 3.3. Although these pseudo depth pairs are useful, they are not highly accurate. Consequently, we found training a model with explicit feature warping using these pseudo depth pairs leads to instability, as shown in Fig. 12.

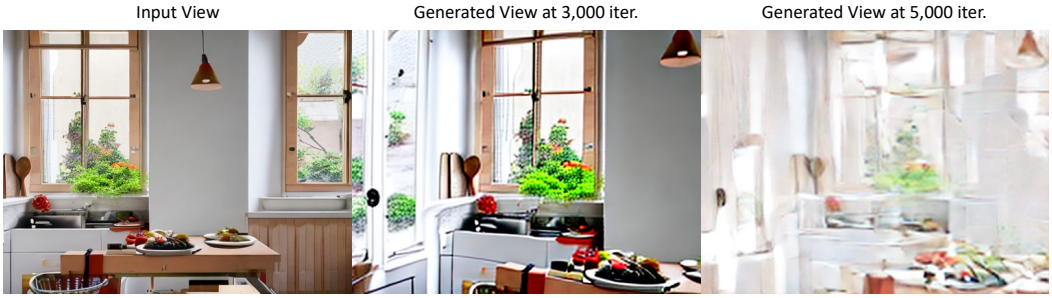

Figure 12: Unstable training of an explicit feature warping model using pseudo depth data.

Additional analysis on viewpoint changes. We analyze how the performance changes as the ratio of pixels invisible from the input view increases due to viewpoint changes. As shown in Fig. 13 and Fig. 14, our method demonstrates the best performance compared to the other methods even as the invisible ratio increases.

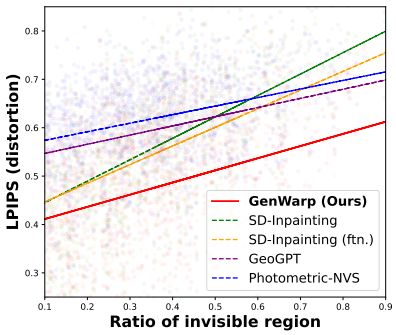

Figure 13: **Comparison of LPIPS with other methods regarding ratio of invisible region.** We measure LPIPS between generated views and GT target views, following GeoGPT [36]'s evaulation protocol.

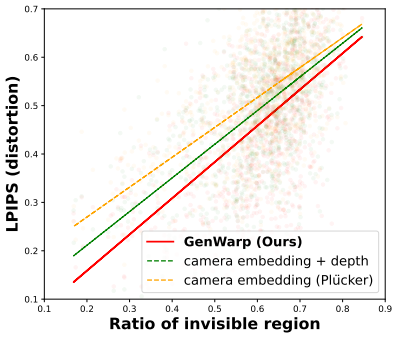

Figure 14: **Additional ablations on view embedding.** We compare our full pipeline with a variant using camera (Plücker) embeddings and another variant using the camera embeddings and depth information.

Comparison on reconstruction-based approaches [11, 50, 28]. We provide comparisons with three additional recent/classical methods (Nerdi [11], PixelNeRF [50], vanilla NeRF [28]) in Fig. 15. For Nerdi [11], due to lack of available codes, we brought curated qualitative results on DTU dataset from their paper. Our result shows non-blurry, clear novel view compared to other methods.

Figure 15: **Comparison with reconstruction-based methods on DTU dataset [18].** Note that our model used here is not trained on DTU dataset.

Analysis on cross-view attention. To verify how the cross-view attentions attend to corresponding points, we have used 1,000 pairs of images to determine (1) how well cross-attention attends to corresponding points, and (2) which is more dominant between self-attention and cross-attention for invisible regions and regions where depth-based correspondence exists. First, Tab. 3 shows the distance between the flow map obtained from depth information and the flow map extracted from the cross-attention. Specifically, we extracted the flow map from the cross-attention layer by argmax operation to see where the model pays the most attention to. It demonstrates that as training progresses, the model learns depth-based matching and warping through the

cross-attention mechanism. On the other hand, the model where the proposed embedding is replaced with the Plücker camera embedding shows relatively worse performance in terms of matching distance.

Secondly, in the Tab. 4, we report which part of the concatenated attention map - the cross-attention part or the self-attention part - is more activated during generation for visible and invisible regions. As exemplified in Fig. 4, it shows the cross-view attention part focuses on regions that can be reliably warped from the input view, while the original self-attention part is more attentive to invisible regions requiring generative priors. Regarding the cross-attention and self-attention for invisible regions, we empirically found that when generating invisible regions, the model also refers to surrounding visible areas through cross-attention, for instance, to generate the invisible left side of a desk, it needs to refer to the visible part of the desk for a plausible novel view.

Models	Average distance
Ours - 2,000 steps	1.36
Ours - 6,000 steps	0.97
Ours - 10,000 steps	0.90
Ours - converged	0.85
Camera embed. - converged	0.98

Table 3: **Matching distance of models over training steps.**

Region	Cross-attn.	Self-attn.
Visible region	0.756	0.244
Invisible region	0.417	0.583

Table 4: **Attention distribution in visible/invisible regions.**

3DGS reconstruction. Our model can be applied to various downstream tasks. For example, given a single image, our model generates 3-4 novel view images, followed by feeding them into fast 3DGS [20] reconstructors such as InstantSplat [12]. Then we can easily obtain a 3DGS scene in a few seconds. Video examples are shown in the project page.

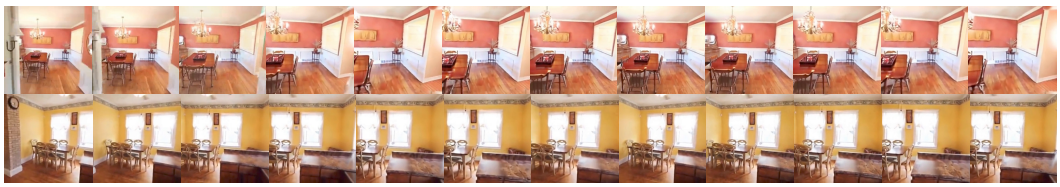

Figure 16: **3DGS reconstruction results.** We show extracted frames in the generated 3DGS videos. Video examples are shown in the project page.

D Limitations

Given extremely distant camera viewpoints where depth-based correspondence has no influence, *i.e.*, beyond the unprojected pixels with the depth map, our model struggles with generating novel views. In these cases, instead of generating a novel view in a single step, the approach of sequentially generating novel views conditioned on pre-generated novel views, similar to other single-shot NVS methods [44, 17, 25, 24], should be taken. As with other works [26, 40, 49] that fine-tune pretrained diffusion models, the quality of the dataset used for fine-tuning affects the model's performance. We believe that more high-quality multi-view datasets will maximize the potential of our model.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The Abstract and Introduction clearly state the motivation and a brief description of the methods used in our paper, along with our contributions. Additionally, at the end of the introduction, we provide a summary of our contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of our method are clearly described in Appendix D.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our paper does not involve theoretical results or their proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In Sec. 3 and Appendix B, we have included the necessary information for reproducibility, such as the network architecture used in our method, which pretrained checkpoint is utilized, which datasets are employed, and what data preprocessing steps are performed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We publicly provide all model weights and code in the project page. The datasets used for training are publicly available.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The implementation details and experimental settings for our method and the baseline methods are described in Sec. 4.1, Sec. 3.3, and Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This task is closer to sampling (generation) within the learned distribution rather than prediction. To be consistent with the experimental analyses of existing works [36, 17], this paper primarily uses distribution distance metrics such as FID and does not include error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the GPU and VRAM used, batch size, and training time in Appendix B.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have faithfully followed the specified NeurIPS Code of Ethics throughout all our experiments.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the societal impacts in Appendix 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We believe that our model, which is a generative model trained on publicly available multi-view datasets, does not have a high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have appropriately cited all the papers corresponding to the datasets we used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not introduce any new assets in this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.