ATTENDIFFCRAFTER: ATTENTION-CRAFTED CONSISTENT NOVEL VIEWS FROM A SINGLE IMAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating consistent novel views from a single image is challenging under large camera motions, where unseen regions must be synthesized while preserving scene structure across frames. Existing approaches face different limitations: autoregressive Transformers tend to accumulate errors over long sequences, whereas diffusion models built on U-Net backbones have limited global interaction and may produce color shifts or implausible content in regions outside the input view. We propose AttenDiffCrafter, a diffusion framework for novel-view synthesis built on attention-centric Transformer architecture. Instead of relying on convolution-dominated feature extraction, the model uses self-attention throughout as the primary mechanisms for spatial and cross-view feature interaction. To better relate different viewpoints, we introduce a learnable query-key cross-attention module that learns view-dependent correspondences between source and target features. Camera motion is represented by relative poses and Plücker coordinates, and temporal attention is used to model the relations along the camera trajectory. Experiments on Matterport3D and RealEstate10K show that AttenDiffCrafter achieved a good overall performance in terms of visual quality, geometric consistency, and temporal coherence.

1 INTRODUCTION

Novel View Synthesis (NVS) aims to generate new views of a scene from known perspectives Avidan & Shashua (1997); Chen & Williams (2023); Laveau & Faugeras (1994). However, synthesizing long-horizon views remains challenging, as the model must *hallucinate* unseen content under large camera motions while preserving consistency.

Most previous methods follow an implicit modeling paradigm to predict the next frame. One line of work employs Transformer-based autoregressive models Rombach et al. (2021); Ren & Wang (2022) to generate future frames sequentially from previous ones, leveraging camera trajectories to learn inter-frame relationships. However, due to error accumulation Liu et al. (2021); Ross et al. (2011) and strong dependence on historical pixels, the generation process often falls into *local optima*, resulting in reduced diversity and limited content richness. Another line of work adopts U-Net-based diffusion models Seo et al. (2024); Tseng et al. (2023a); Yu et al. (2023); Müller et al. (2024); Yu et al. (2024); Zhang et al. (2025); Zheng & Li (2026), which samples from learned 3D scene priors to produce more diverse video content. However, their locally focused convolutional architectures (e.g., ResNet blocks He et al. (2016)) and sparsely applied attention typically restricted to low-resolution bottleneck layers limit the propagation of global context. As a result, high-resolution features receive little global information and remain dominated by local convolutions, often leading to global tone inconsistencies and semantically implausible "unknown regions" in unobserved areas. In particular, ViewCrafter Yu et al. (2024) tends to change the overall scene color style, while SceneSplatter Zhang et al. (2025) struggles under large viewpoint changes.

To overcome these limitations, we propose **AttenDiffCrafter**, a novel diffusion-based generative framework built upon the Diffusion Transformer (DiT) Peebles & Xie (2023). Our method adopts an **self-attention-based diffusion architecture** inspired by Vision Transformer (ViT) Dosovitskiy et al. (2021), reducing reliance on convolutional feature extraction and enabling stronger global semantic modeling. This helps mitigate the tone-shift and "unknown block" artifacts commonly observed in convolution-based counterparts (see Figure 1). Notably, our framework requires **no pretrained**

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

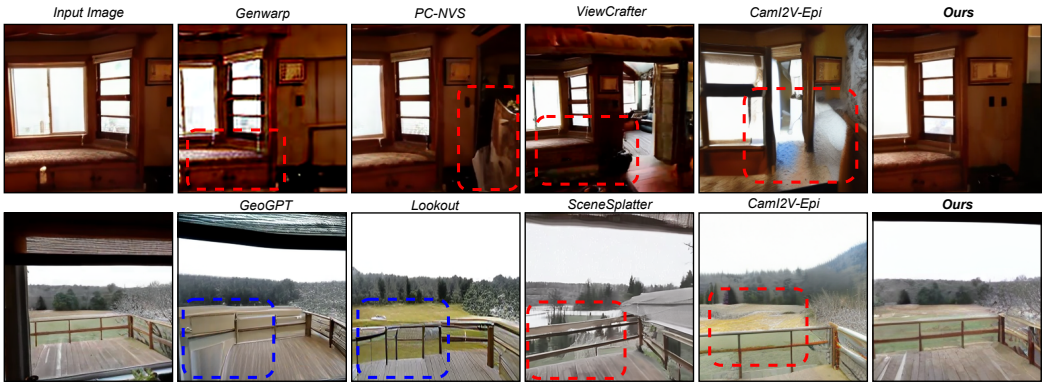

Figure 1: The U-Net diffusion model suffers from "color shift" and "unknown regions", while Transformer-based models face "local optima" (red and blue boxes).

video diffusion models, in contrast to existing approaches such as ViewCrafter and SceneSplatter that rely on strong video priors. From a modeling perspective, AttenDiffCrafter introduces the attention-centric NVS architecture into the diffusion-based trajectory generation method, and further models the viewpoint-dependent cross-view correspondence through LQCA (**learnable query–key cross-attention**). When combined with camera Plücker encoding, this mechanism significantly enhances inter-frame geometric consistency by capturing viewpoint-dependent correspondences between source and target views. From a control perspective, we encode the camera trajectory as **multi-scale pose embeddings** within the framework, allowing temporal attention to effectively model long-range frame dependencies and ensure coherent view transitions across the generated sequences.

We conduct extensive experiments on two widely used 3D scene datasets:**RealEstate10K** Zhou et al. (2018) and **Matterport3D** Chang et al. (2017). The results demonstrate that AttenDiffCrafter can synthesize coherent and high-quality video sequences under both large and small camera motions (see Figure 2). It achieves a good overall performance in terms of visual quality, geometric consistency, and temporal coherence in short-term, medium-term, and long-term future frame generation experiments. Our main contributions are summarized as follows:

1.Pure self-attention diffusion architecture. AttenDiffCrafter processes tokenized patch embeddings entirely without convolutions, enabling global information exchange at every layer to eliminate color shifts and "unknown region" artifacts under large camera motions.

2.Learnable query-key cross-attention (LQCA). A novel cross-attention module that implicitly learns view-dependent geometric correspondences without explicit 3D priors or handcrafted constraints. Combined with Plücker embedding and temporal attention, it ensures strong cross-frame geometric consistency and smooth view transitions.

3.Superior performance in novel view synthesis. Our method achieves strong results on both **RealEstate10K** and **Matterport3D**, performing well across short-, mid-, and long-term novel view synthesis tasks, with particularly notable improvements on the challenging Matterport3D dataset.

2 Related Work

2.1 Transformer-based Models for NVS.

With the advancement of single-image-based novel view synthesis (NVS) techniques, researchers begin exploring the animation of static photographs Wiles et al. (2020); Tancik et al. (2022); Liu et al. (2023). Wiles et al. Wiles et al. (2020) use 3D point clouds as an intermediate representation for view synthesis, while Lai et al. Liu et al. (2023) propose a video autoencoder. However, these approaches struggle to extrapolate to regions far from the source view. Inspired by the success of Transformer models in language modeling Huang et al. (2024); Hu et al. (2021); Lai et al. (2021), researchers apply them to image and video generation tasks Devlin et al. (2019); Radford et al. (2018); Vaswani et al. (2017). GeoGPT Rombach et al. (2021) introduces an autoregressive Transformer-based

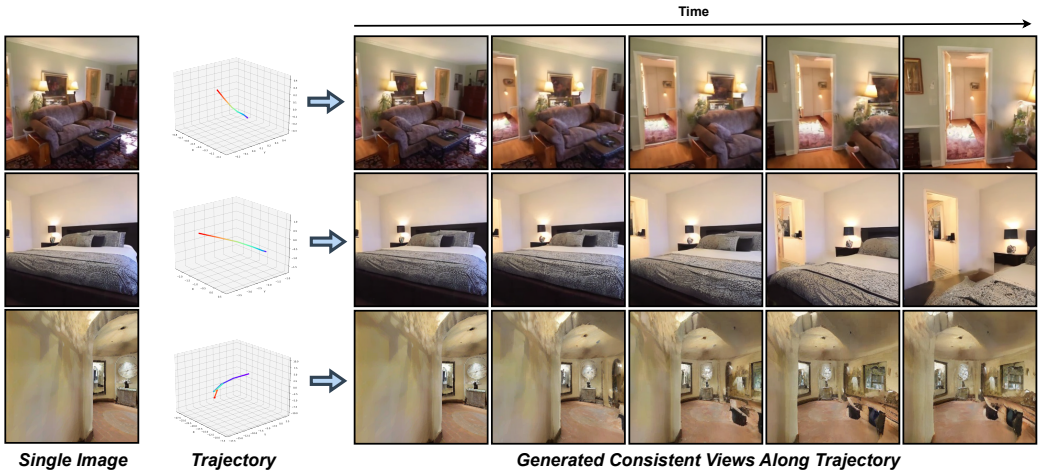

Figure 2: Given a single image, AttenDiffCrafter generates consistent novel views following a desired camera trajectory, while plausibly extrapolating unseen regions. Examples from RealEstate10K Zhou et al. (2018) (top two rows) and Matterport3D Chang et al. (2017) (bottom row) test sets demonstrate that our model can handle large or small camera changes and challenging perspectives.

framework that implicitly models geometric transformations via global attention without requiring explicit 3D priors. While GeoGPT produces diverse novel views, it often lacks consistency along continuous camera trajectories. LookOut Ren & Wang (2022) extends this approach by conditioning each frame on the previous two frames, thereby improving temporal coherence. LVSM Jin et al. (2025) have demonstrated that Transformer-based architectures with minimal 3D inductive bias can effectively perform novel-view synthesis.However, these methods focus on feed-forward novel-view reconstruction, while our diffusion framework progressively synthesizes unseen regions along long camera trajectories.

2.2 Diffusion-based Models for NVS.

With the impressive success of diffusion models in image synthesis Ho et al. (2020); Karras et al. (2024); Rombach et al. (2022), particularly in video generation Ho et al. (2022b); Singer et al. (2022); Voleti et al. (2022) and 3D shape modeling Luo & Hu (2021); Vahdat et al. (2022); Zhou et al. (2021), an increasing number of studies explore incorporating diffusion processes into NVS tasks. Tseng et al. Tseng et al. (2023b) propose a pose-guided diffusion model that incorporates a cross-attention mechanism constrained by epipolar geometry to guide the diffusion process. However, its cascaded architecture Ho et al. (2022a) leads to high computational cost, unstable gradient propagation, and susceptibility to local minima. Elata et al. Elata et al. (2024) introduce a pixel-space diffusion model that avoids high-frequency detail loss from autoencoder reconstruction, but suffers from high complexity and limited generalization. PC-NVS Yu et al. (2023) adopts a dual-stream latent diffusion architecture with camera-pose-guided cross-attention. GenWarp Seo et al. (2024) fine-tunes two pre-trained text-to-image latent diffusion models Rombach et al. (2022), replacing the text embeddings with CLIP image embeddings Radford et al. (2021) and aligning 2D coordinate embeddings via monocular depth estimation. ViewCrafter Yu et al. (2024) proposes a point-conditioned video diffusion framework that iteratively refines camera trajectories, but it tends to alter the scene's color style, resulting in visible color shifts. SEVAZhou et al. (2025) eliminates explicit 3D representations for novel view generation but incurs high computational costs and scale ambiguity. SceneSplatter Zhang et al. (2025) employs a momentum-based strategy to balance generative priors with existing scene cues, yet it performs well only under small viewpoint variations. In contrast, CamI2V-EpiZheng & Li (2026) integrates epipolar geometric priors into the block selection stage of video block sparse attention, yet its current implementation is limited to a binary hard selection strategy.

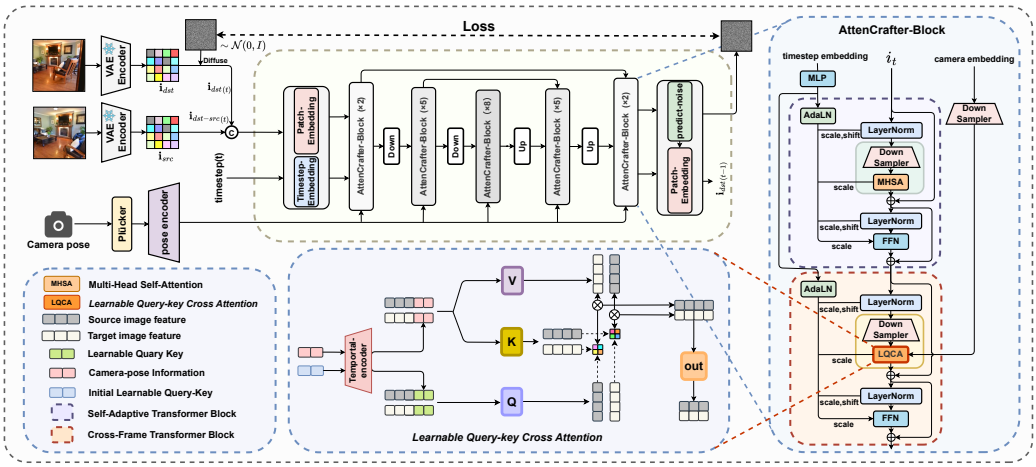

Figure 3: **Method overview.** We propose**AttenDiffCrafter**, a novel view synthesis framework with a fully Transformer-based NVS framework that enables global information exchange and structural consistency under wide camera motions. Its learnable query–key cross-attention (LQCA) captures viewpoint-dependent interactions and geometric correspondences, improving spatial-temporal coherence and view consistency.

3 METHOD

In this section, we present **AttenDiffCrafter**, a diffusion-based framework for generating geometrically consistent novel-view sequences from a single image(Fig. 3). To overcome the local optimum of autoregressive methods and the limited locality of U-Net diffusion models, we adopt a fully transformer-based backbone with strong global modeling capabilities for long sequences and direct sampling from 3D scene priors, alleviating the "unknown region" problem

3.1 OVERALL ARCHITECTURE

Two-stream architecture. Our model adopts a two-stream architecture Seo et al. (2024); Elata et al. (2024) with shared weights between the two branches. One branch takes $I_{\text{dst}} \in \mathbb{R}^{b\times H\times W\times C}$ as the target image, while the other uses $I_{\text{src}} \in \mathbb{R}^{b\times H\times W\times C}$ as part of the conditioning input. Both inputs are first encoded into latent representations i_{dst} and i_{src} using AutoencoderKL Rombach et al. (2022). To preserve maximal scene information, we do not apply diffusion noise to i_{src}. The two-stream network extracts multi-scale features from both the noisy i_{dst} and the clean i_{src}, and fuses them via an interleaved concatenation operation:

$$i_{\text{dst-src}} = \text{Concat}(i_{\text{dst}} + noise,\, i_{\text{src}}) \in \mathbb{R}^{2b\times h\times w\times c} \tag{1}$$

This fused representation provides the model with paired geometric cues and appearance information at each layer of the network.

Tokenization and Attention Processing. The fused representation $i_{\text{dst-src}}$ is projected into the token space through a patch embedding layer, facilitating efficient attention computation. Formally, we define:

$$z_0 = \text{PatchEmbedding}(i_{\text{dst-src}}) \in \mathbb{R}^{b\times N\times d} \tag{2}$$

where $N = \frac{h\times w}{p^2}$ is the number of tokens, p denotes the patch size, and d is the embedding dimension. These tokens are then processed by a stack of AttenDiffCrafter-Blocks introduced in Sec. 3.2, progressively refine spatial–temporal correspondences under camera-guided conditioning:

$$z_{l+1} = \text{AttenDiffCrafterBlock}_l(z_l, \dots) \tag{3}$$

Each block integrates multi-scale attention and learnable query–key modulation to enhance geometric alignment and view-consistent feature fusion across tokens.

3.2 ATTENDIFFCRAFTER-BLOCK FOR NOVEL VIEW SYNTHESIS

We introduce the **AttenDiffCrafter Block** (Fig 3), composed of two components, each serving a distinct purpose in ensuring both spatial and temporal coherence in novel-view synthesis.

Self-Adaptive Transformer Block. This block extends standard multi-head self-attention by integrating timestep-conditioned adaptive layer normalization (AdaLN) and a DownSampler module. It dynamically modulates feature representations according to temporal context, preserving global structural consistency and capturing long-range dependencies within a single frame.

Cross-Frame Transformer Block. This block uses cross-attention to integrate multi-frame information. We introduce a learnable camera-aware query-key mechanism (LQCA) guided by camera pose embeddings to align features across views. DownSampler reduces computation, while AdaLN enhances cross-frame feature integration, improving temporal consistency and scene geometry under large viewpoint variations.

To reduce attention computation costs while maintaining performance, we incorporate the DownSampler module proposed in Tian et al. (2024) and replace standard convolution with pixel shuffle operations. This design effectively compresses the token length in long sequences while preserving the structural integrity of the original token representations.

Camera information is then used to provide geometric priors that guide the cross-attention module, which models cross-frame relationships and enables plausible view transformations. We detail this module in Sec. 3.3.

3.3 CAMERA POSE CONDITIONING

Our objective is to generate a sequence of spatially and temporally consistent frames $\{I_i\}_{i=1}^{n}$ from a single input image I_0, guided by corresponding camera poses $\{K_i, R_i, T_i\}_{i=1}^{n}$, where K_i, R_i, and T_i denote the intrinsic, rotation, and translation parameters, respectively. Each view's full camera pose can be represented by a homogeneous transformation matrix $RT_i \in \mathbb{R}^{4\times4}$ that combines rotation and translation. To synthesize a target view I_{dst} from a source view I_{src}, we compute the **relative camera pose** as:

$$RT_{\text{ref}} = RT_{\text{dst}} \cdot RT_{\text{src}}^{-1} \tag{4}$$

which describes the transformation from the source to the target camera frame. See Appendix A for details.

In our model, $RT_{s\to d}$ and $RT_{d\to s}$ represent bidirectional relative poses between I_{src} and I_{dst}, serving as the reference transformation RT_{ref} required by the two-stream architecture. Together with the intrinsic matrix K, they form the camera condition $c_{\text{cam}} = [RT_{\text{ref}} \mid K]$ used in both training and inference.

Previous works Rombach et al. (2021); Yu et al. (2023); Ren & Wang (2022) typically flatten RT_{ref} for MLP processing, which may cause inconsistency due to unbounded translations, or use ray-based encodings Vaswani et al. (2017) that increase geometric complexity. In contrast, we explicitly model the bidirectional relationship between $RT_{s\to d}$ and $RT_{d\to s}$ through a temporal-attention mechanism, enabling smoother view transitions. The effectiveness of the Plücker embedding and temporal-attention module is verified in Sec. 4.4.

Following the design in CameraCtrl He et al. (2024), we adopt a Plücker coordinate embedding for the camera condition, defined as:

$$d_{u,v} = R_{\text{ref}}^{-1}[u, v, 1]^{\mathsf{T}} + T_{ref}, p_{u,v} = (o \times d_{u,v},\ d_{u,v}) \in \mathbb{R}^6 \tag{5}$$

where $o \in \mathbb{R}^3$ is the camera origin in world coordinates (equivalent to T_{ref}), and $d_{u,v} \in \mathbb{R}^3$ represents the direction from the camera center to the pixel (u, v) in world space. This formulation yields a Plücker embedding $P \in \mathbb{R}^{2\times6\times H\times W}$. A learnable pose encoder is then used to transform this into a multi-scale camera embedding $Ce \in \mathbb{R}^{2\times c\times h\times w}$, which fits the input requirements of our model.

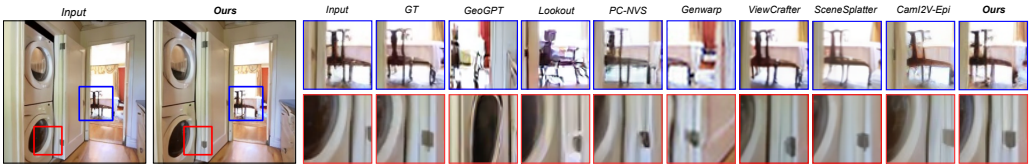

Figure 4: **Qualitative result of frame generation quality.** We provide view synthesis results of a representative frame along the generated trajectory for different methods. The shown patches are uniformly cropped from the same spatial region, aligned with the corresponding location in the second frame generated by our method.

3.4 Cross-frame Modeling for View Synthesis

To effectively model the strong correspondence between the source and target views, we propose a **learnable query-key cross-attention mechanism** (in Fig. 3). While this entails higher computational complexity inherent to the transformer-based design, we employ effective strategies to mitigate the cost.

DownSampler for Efficient Attention. Given a feature map $X \in \mathbb{R}^{C\times H\times W}$, the quadratic attention cost $\mathcal{O}((HW)^2)$ becomes prohibitive at high resolutions. We introduce a lightweight *DownSampler* that reduces the spatial resolution by a factor of s before cross-attention, lowering the token count to $\frac{HW}{s^2}$ and the complexity to $\mathcal{O}\Big(\big(\frac{HW}{s^2}\big)^2\Big) = \mathcal{O}\Big(\frac{(HW)^2}{s^4}\Big)$. The module first applies a depthwise convolution$\tilde{X}_c = W_c * X_c$ And implemented as $conv(\cdot, \text{groups} = C)$, optionally followed by a residual shortcut $X' = \tilde{X} + X$. A deterministic folding operation then rearranges spatial neighborhoods into channels:

$$Y = \text{Fold}(X') \in \mathbb{R}^{s_2\times(H/s)\times(W/s)\times C} \tag{6}$$

This design introduces negligible computational overhead while substantially reducing attention memory usage (from 19.07 GB to 13.64 GB in our model), enabling efficient cross-attention at high resolutions.The diffusion backbone is predominantly attention-based. A lightweight depth-wise convolution is retained only in the DownSampler for spatial reduction and computational efficiency.

Cross-Attention with Learnable Query-Keys. We first extract image features $F_{\text{src}}, F_{\text{dst}} \in \mathbb{R}^{c\times h\times w}$ and camera embeddings $Ce_{s\to d}, Ce_{d\to s} \in \mathbb{R}^{c\times h\times w}$ from the source and target views. These are passed through a **DownSampler** to obtain compact representations $f_{\text{src}}, f_{\text{dst}}, ce_{s\to d}, ce_{d\to s} \in \mathbb{R}^{n^2\times c\times \frac{hw}{n^2}}$, which reduce token length while preserving spatial structure. To allow the camera embeddings to guide the generation of the next frame, we concatenate them with the corresponding image features along the channel dimension, as they share the same shape. The combined representations are then used to compute the queries, keys, and values for cross-attention:

$$\begin{aligned} q_{\text{src}} &= L([f_{\text{src}}, \quad lq_{\text{src}}]), \quad q_{\text{dst}} = L([f_{\text{dst}}, \quad lq_{\text{dst}}]) \\ k_{\text{dst}} &= L([f_{\text{dst}}, ce_{s\to d}]), \quad k_{\text{src}} = L([f_{\text{src}}, ce_{d\to s}]) \\ v_{\text{dst}} &= L([f_{\text{dst}}, ce_{s\to d}]), \quad v_{\text{src}} = L([f_{\text{src}}, ce_{d\to s}]) \end{aligned} \tag{7}$$

Here, lq_{src} and lq_{dst} are the proposed **learnable query-keys**, initialized as $I_{4\times 4}$ and transformed using the same operations applied to RT_{ref}. Notably, they do not contain explicit camera pose information. Instead, during training, they learn view-to-view priors that align them with the camera embeddings in a shared space. This introduces beneficial stochasticity and improves generalization. The full pipeline is illustrated in Fig. 3.

3.5 Training and Inference

We train AttenDiffCrafter end-to-end with a unified diffusion-based loss, enabling the model to learn spatial, temporal, and geometric consistency from a single image. During inference, we adopt the DDIM sampling strategy Song et al. (2020) for efficient and consistent novel-view generation.

Training objective. Our model is fully end-to-end, allowing training with a single loss function. We adopt the standard DDPM objective Ho et al. (2020) and further incorporate the improved DDPM loss Nichol & Dhariwal (2021), which introduces an additional variational lower bound term $\mathcal{L}_{\text{vlb}}$. We use mean squared error (MSE) as the core training objective, formulated as:

$$\mathcal{L}_{\text{AttenDiffCrafter}} = \mathbb{E}_{i_{\text{dst}},t}\left[\left\|\epsilon_t - \epsilon_\theta(i_{\text{src}}, i_{\text{dst}_t}, RT_{s\to d}, RT_{d\to s}, RT_{lq}, K, t)\right\|_2^2\right] + \mathcal{L}_{\text{vlb}} \tag{8}$$

Here ϵ_t is the noise added to the target image at timestep t, and $\epsilon_\theta(\cdot)$ is the denoising network predicting this noise. RT_{lq} denotes the initialized Learnable Query-Key, and K is the camera intrinsics. The $\mathcal{L}_{\text{vlb}}$ term is derived from the log-variance prediction head of $\epsilon_\theta(\cdot)$.

Inference stage. During inference, we assume a given reference image i_{src} and its associated camera sequence $C = [RT_1, RT_2, \ldots, RT_n, K]$, which are used to construct the condition c. We adopt the DDIM sampling strategy Song et al. (2020) to generate a target video sequence from our model. Starting from $i_{\text{dst}_t} \sim \mathcal{N}(0, I)$, the generation of each frame proceeds step by step as follows:

$$\mu_\theta(i_{\text{dst}_t}, t) = \frac{1}{\sqrt{\alpha_t}}\left(i_{\text{dst}_t} - \left(\frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\right)\epsilon_\theta(i_{\text{dst}_t}, c)\right)$$

$$i_{\text{dst}_{t-1}} = \mu_\theta(i_{\text{dst}_t}, t) + \Sigma_\theta(i_{\text{dst}_t}, c)\cdot\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I) \tag{9}$$

Here, i_{dst_t} denotes the intermediate denoising state at timestep t, and $i_{\text{dst}_{t-1}}$ is the corresponding denoised result. The predicted variance $\Sigma_\theta(i_{\text{dst}_t}, c)$ controls the noise scale applied to ϵ_t, allowing for flexible stochastic sampling and improved diversity in generated sequences.

4 Experiments

In this section, we evaluate our method's performance on the task of consistent novel view synthesis from a single reference image and conduct ablation studies to verify the efficacy of our design.

4.1 Experimental Setup

We train our model on two multi-view datasets focusing primarily on indoor scenes: the real-world RealEstate10K (Re10K) Zhou et al. (2018) and the synthetic Matterport3D (MP3D) Chang et al. (2017). We compare our method with two categories of representative approaches: Transformer-based methods including GeoGPT Rombach et al. (2021) and LookOut Ren & Wang (2022), and recent U-Net-based diffusion models such as PC-NVS Yu et al. (2023), GenWarp Seo et al. (2024),CamI2V-Epi Zheng & Li (2026),ViewCrafter Yu et al. (2024), and SceneSplatter Zhang et al. (2025). For evaluation, we follow the metric protocol used in ViewCrafter Yu et al. (2024) and SceneSplatter Zhang et al. (2025), reporting PSNR, SSIM, LPIPS, FID, and FVD under short-, mid-, and long-term prediction settings. Our method delivers competitive results and surpasses both categories of baselines. We evaluate view synthesis in short-, mid-, and long-term scenarios by generating 20-frame videos for each test sequence, following prior works. The first 3 and 6 frames represent short- and mid-term settings, respectively.

To evaluate the geometric consistency of generated views, we follow a pose estimation procedure similar to prior work Yu et al. (2024). For each generated frame, we estimate its camera pose using Dust3r Wang et al. (2024). All poses are transformed to be *relative* to the first frame, and translation vectors are *normalized* using the furthest-frame distance. We compute two metrics to quantify pose accuracy: rotation distance (Rdist) and translation distance (Tdist), see more details in Appendix F.

We implement AttenDiffCrafter using PyTorch platform. The model is optimized with AdamW using a learning rate of 1×10^{-4} and no weight decay. AdamW decouples weight decay from gradient updates, which improves training stability. We employ a linear noise scheduler $\{\beta_t\}_{t=1}^{T}$, where $\beta_1 = 1 \times 10^{-4}$ and $\beta_T = 2 \times 10^{-2}$. During training, the total number of diffusion time steps T is set to 1,000. For inference, we use the DDIM Song et al. (2020) sampling strategy with 250 steps to generate high-quality sequences. For more details, please refer to Appendix F.

Table 1: **Quantitative comparisons.** We report the average PSNR (↑), SSIM (↑), and LPIPS (↓) scores between the generated results and the ground truth over the short-term, middle-term, and long-term settings on the test sets of the MP3D and RealEstate10K datasets. We report frame consistency F-C (↑), computed using CLIP ViT-L/14 Radford et al. (2021), to evaluate long-term novel view synthesis.

	Method	Short-term			Middle-term			Long-term					
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	F-C↑	R_{dis}↓	T_{dis}↓
MP3D	GeoGPT	15.446	0.507	3.024	14.309	0.474	3.245	12.165	0.406	3.598	0.842	24.609	0.493
	LookOut	16.245	0.543	2.782	15.000	0.517	2.980	12.939	0.477	3.353	0.885	9.336	0.438
	PC-NVS	16.174	0.550	2.937	15.354	0.530	3.113	14.169	0.500	3.366	**0.909**	13.147	0.432
	Ours	**18.887**	**0.638**	**2.234**	**17.393**	**0.589**	**2.605**	**15.415**	**0.518**	**3.156**	0.899	**8.941**	**0.416**
Re10K	GeoGPT	15.730	0.470	2.488	14.492	0.427	2.748	11.791	0.340	3.284	0.879	8.156	**0.134**
	LookOut	15.090	0.472	2.534	14.056	0.442	2.805	11.856	0.383	3.297	0.882	11.001	0.264
	PC-NVS	18.775	0.613	1.968	17.300	0.565	2.249	14.374	0.462	2.889	0.902	4.387	0.173
	GenWarp	18.435	0.563	2.214	16.025	0.490	3.660	12.657	0.397	3.287	0.913	8.153	0.577
	CamI2V-Epi	16.115	0.507	2.416	14.772	0.464	2.741	12.253	0.379	3.355	0.864	9.351	0.242
	ViewCrafter	16.713	0.559	2.024	15.305	0.509	2.311	12.421	0.401	2.929	**0.923**	5.254	0.162
	SceneSplatter	17.734	0.614	2.263	16.783	0.577	2.490	14.152	0.480	3.005	0.912	4.940	0.153
	Ours	**19.441**	**0.635**	**1.780**	**17.916**	**0.583**	**2.106**	**14.819**	**0.481**	**2.816**	0.893	**3.962**	0.169

4.2 Qualitative Results

Qualitative comparison of frame generation quality. As shown in Fig. 4, we compare next-frame predictions from different methods using two patches cropped at the same spatial locations. Our method produces more coherent and detailed results, demonstrating its ability to effectively leverage camera information for high-quality frame generation.

Qualitative comparison of extrapolation capabilities. Appendix Fig. 5 compares long-horizon view generation from a single image. As the camera moves right to left, the reflected door in the mirror should naturally appear. GeoGPT Rombach et al. (2021) and LookOut Ren & Wang (2022) remain static under large motions, revealing limited extrapolation. PC-NVS Yu et al. (2023) produces unnatural blocks, and GenWarp Seo et al. (2024) yields warped artifacts with color inconsistency. ViewCrafter Yu et al. (2024) introduces noticeable color shifts due to point-conditioned diffusion, CamI2V-EpiZheng & Li (2026) generates unsatisfactory results, while SceneSplatter Zhang et al. (2025) fails to maintain consistency under larger motions. In contrast, AttenDiffCrafter plausibly completes the unseen door and generates coherent, consistent sequences without any 3D priors.

Table 2: FID and FVD scores on short-term video generation. Lower is better.

Method	MP3D		Re10K	
	FID↓	FVD↓	FID↓	FVD↓
GeoGPT	218.78	1204.15	113.66	942.98
LookOut	213.49	926.15	80.45	689.27
PC-NVS	200.21	696.49	93.23	622.41
GenWarp	N/A	N/A	111.12	845.66
CamI2V-epi	N/A	N/A	106.72	849.47
ViewCrafter	N/A	N/A	**65.11**	794.21
SceneSplatter	N/A	N/A	175.06	1243.31
Ours	**133.39**	**484.93**	68.78	**433.52**

Table 3: **Ablation on LQCA.** Comparison of attention variants on MP3D at 100k, 200k, and 300k steps("N/A": these methods were not evaluated on this dataset).

Method	100k		200k		300k	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
EA	16.689	2.856	17.076	2.707	17.218	2.651
E4A	17.309	2.783	17.465	2.642	17.696	2.553
Ours	**18.152**	**2.424**	**18.312**	**2.349**	**18.557**	**2.320**

4.3 Quantitative Results

We train all methods separately on the MP3D and RealEstate10K datasets. As shown in Table 1, GeoGPT Rombach et al. (2021) produces low-quality results, particularly in long-term generation, while LookOut Ren & Wang (2022) provides slightly improved fidelity but still struggles with temporal consistency. PC-NVS Yu et al. (2023) maintains high image fidelity but often exhibits structural artifacts under large camera motions. GenWarp Seo et al. (2024) introduces noticeable distortions in long-horizon synthesis. ViewCrafter Yu et al. (2024) suffers from color shifts due to its

point-conditioned diffusion process, and SceneSplatter Zhang et al. (2025) generates results often appear whitened and desaturated. It is worth noting that although AttenDiffCrafter does not achieve the highest F-Consistency on the long-term setting, it consistently provides a stronger balance of appearance fidelity and geometric consistency, as evidenced by the PSNR, SSIM, LPIPS, Rdist, and Tdist results.

As shown in Table 2, AttenDiffCrafter also obtains notably improved FID and FVD scores on both MP3D and RealEstate10K datasets for short-term, significantly outperforming prior Transformer- and UNet-based methods. AttenDiffCrafter achieves competitive visual quality and substantially improved FVD scores on the short-term setting. On RealEstate10K, although ViewCrafter obtains a slightly lower FID, AttenDiffCrafter achieves a substantially lower FVD (433.52 vs. 794.21), indicating stronger temporal coherence. These results demonstrate that AttenDiffCrafter effectively generates realistic and temporally coherent video sequences from a single image.

4.4 Ablation Study

We further study the effect of LQCA through two variants: (1) **Epipolar Attention (EA)**: standard epipolar-constrained attention along a thin line; (2) **Epipolar Aggregate4 Attention (E4A)**: aggregates tokens over 4×4 regions to simulate a thicker epipolar line. Table 3 shows that both variants underperform our full model, indicating that implicit geometric priors through learnable attention outperform explicit geometric constraints. More details are provided in Appendix E. These results highlight the effectiveness of LQCA with implicit geometric priors.

To evaluate the impact of plücker encoding of RT_{ref} and temporal attention introduced in the pose encoder, we conduct ablation studies. These experiments analyze their individual contributions as well as their combined effect, providing insights into their roles and complementarity within the model. A comparison in Table 4 shows that both Plücker encoding and Temporal Attention contribute to model performance individually. Combining them yields the best results, indicating a synergistic effect. Plücker encoding enhances spatial reasoning, and Temporal Attention improves temporal coherence, jointly enabling more accurate view synthesis.

Table 4: **Ablation on RT Process.** We report the short-term FVD results on the MP3D dataset after training for 200k steps.

Method	plücker	temporal attention	FVD↓
Model1	✗	✗	582.02
Model2	✓	✗	547.82
Model3	✗	✓	552.80
Ours	✓	✓	**507.47**

Table 5: **Block Configuration.** First-frame performance comparison of AttenDiffCrafter block distributions at 200k training steps

Configuration	Blocks	PSNR ↑	LPIPS ↓
Shallow	1-1-1-1-1	19.604	2.013
Balanced	1-3-4-3-1	19.844	2.013
Deep	4-8-10-8-4	19.979	1.989
Ours	**2-5-8-5-2**	**20.452**	**1.939**

To assess the effect of block distribution across stages, we vary the number of blocks per stage while keeping the total model size roughly constant by adjusting linear dimensions. We compare four configurations: *Shallow* (1-1-1-1-1), *Balanced* (1-3-4-3-1), *Deep* (4-8-10-8-4), and our default (2-5-8-5-2). All models are trained for 200k steps and evaluated on the first generated frame using PSNR and LPIPS.As shown in Table 5, increasing depth generally improves performance, with *Deep* performing better than *Shallow* and *Balanced*. However, our default design achieves the highest PSNR (20.452) and lowest LPIPS (1.939), indicating that a moderate block increase in the middle stages provides an optimal balance between model capacity and efficiency.

5 Conclusion

We present **AttenDiffCrafter**, an attention-driven diffusion framework that generates geometrically consistent novel-view sequences from a single image and a camera trajectory while preserving scene appearance. Its Transformer backbone enables global feature exchange under large camera motions, while learnable query–key cross-attention strengthens viewpoint-dependent geometric reasoning. Experiments on Matterport3D and RealEstate10K confirm its effectiveness relative to prior methods.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code debugging and refinement, manuscript formatting, grammar and readability improvement, reference and citation checking, and figure visualization enhancement. The research ideas, proposed method, experimental design, data analysis, and scientific conclusions were developed by the authors. All AI-assisted code, text, and visual materials were carefully reviewed and verified by the authors before inclusion in the manuscript. AI tools were not used as a substitute for scientific judgment. The authors take full responsibility for all the claims and conclusions of this work.

REPRODUCIBILITY STATEMENT

We describe the AttenDiffCrafter architecture, camera pose conditioning, cross-frame modeling, and training objective in Section 3. Section 4 specifies the datasets, evaluation protocols and metrics, baseline methods, training and inference settings, and ablation studies used in our experiments. Further details required for reproduction are provided in the appendix: Appendix A gives the pose and geometric formulations; Appendix E documents the attention variants; Appendix F reports data preprocessing, metric definitions, model configuration, hardware, optimization, and sampling settings; and Appendix D provides supplementary qualitative comparisons. Together, these descriptions specify the model, training procedure, evaluation setup, and analyses needed to reproduce the reported results.

REFERENCES

Shai Avidan and Amnon Shashua. Novel view synthesis in tensor space. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1034–1040. IEEE, 1997.

Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. *arXiv preprint arXiv:2412.18597*, 2024.

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.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 423–432. 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. URL https://arxiv.org/abs/2010.11929.

Noam Elata, Bahjat Kawar, Yaron Ostrovsky-Berman, Miriam Farber, and Ron Sokolovsky. Novel view synthesis with pixel-space diffusion models. *arXiv preprint arXiv:2411.07765*, 2024.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022b.

Ronghang Hu, Nikhila Ravi, Alexander C Berg, and Deepak Pathak. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12528–12537, 2021.

Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9784–9794, 2024.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. 2025. URL `https://arxiv.org/abs/2410.17242`.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24174–24184, 2024.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022.

Zihang Lai, Sifei Liu, Alexei A Efros, and Xiaolong Wang. Video autoencoder: self-supervised disentanglement of static 3d structure and motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9730–9740, 2021.

Stephane Laveau and Olivier D Faugeras. 3-d scene representation as a collection of images. In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pp. 689–691. IEEE, 1994.

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*, pp. 14458–14467, 2021.

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*, pp. 9298–9309, 2023.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2837–2845, 2021.

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*, pp. 10258–10268, 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Tianhao Qi, Jianlong Yuan, Wanquan Feng, Shancheng Fang, Jiawei Liu, SiYu Zhou, Qian He, Hongtao Xie, and Yongdong Zhang. Dit: Dual mask-based diffusion transformer for multi-scene long video generation. *arXiv preprint arXiv:2503.19881*, 2025.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training.(2018), 2018.

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*, pp. 8748–8763. PmLR, 2021.

Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing a consistent long-term 3d scene video from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3563–3573, 2022.

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*, pp. 14356–14366, 2021.

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*, pp. 10684–10695, 2022.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

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.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9339–9347, 2019.

Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

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

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8248–8258, 2022.

Yuchuan Tian, Zhijun Tu, Hanting Chen, Jie Hu, Chao Xu, and Yunhe Wang. U-dits: Downsample tokens in u-shaped diffusion transformers. *arXiv preprint arXiv:2405.02730*, 2024.

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*, pp. 16773–16783, 2023a.

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*, pp. 16773–16783, 2023b.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022.

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.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Masked conditional video diffusion for prediction. *Generation, and Interpolation*, 2022.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

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*, pp. 7467–7477, 2020.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7623–7633, 2023.

Jason J Yu, Fereshteh Forghani, Konstantinos G Derpanis, and Marcus A Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7094–7104, 2023.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

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*, pp. 586–595, 2018.

Shengjun Zhang, Jinzhao Li, Xin Fei, Hao Liu, and Yueqi Duan. Scene splatter: Momentum 3d scene generation from single image with video diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6089–6098, 2025.

Guangcong Zheng and Xi Li. Cami2v-epipolar: Epipolar-constrained block sparse attention for camera-controlled image-to-video diffusion model. *IEEE Signal Processing Letters*, 2026.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025.

Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5826–5835, 2021.

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.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1219–1229, 2023.

A CAMERA POSE DERIVATION

In both the MP3D (Matterport3D) Chang et al. (2017)and RealEstate10K Zhou et al. (2018) datasets, each viewpoint image is associated with a set of extrinsic parameters, namely a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{T}$. Together, they describe the camera's orientation and position in the 3D world coordinate system, serving as the key parameters for transforming coordinates from the world coordinate system to the camera coordinate system.

Assuming a point P has coordinates (x_w, y_w, z_w) in the world coordinate system, we can obtain its coordinates in the camera coordinate system as follows:

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \mathbf{R} \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} + \mathbf{T}$$

By introducing homogeneous coordinates, the rotation matrix $\mathbf{R}$ and translation vector $\mathbf{T}$ can be combined into a single 4×4 extrinsic matrix $\mathbf{RT} := [\mathbf{R} \mid \mathbf{T}]$:

$$\begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ \mathbf{0}^\top & 1 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}$$

where $\mathbf{R} \in \mathbb{R}^{3\times3}$ is the rotation matrix, $\mathbf{T} \in \mathbb{R}^{3\times1}$ is the translation vector, and the full extrinsic matrix is in $\mathbb{R}^{4\times4}$.

To compute the relative transformation from the source camera coordinate system $(\mathbf{R}_{\text{src}}, \mathbf{T}_{\text{src}})$ to the target camera coordinate system $(\mathbf{R}_{\text{dst}}, \mathbf{T}_{\text{dst}})$, we first express a 3D point in world coordinates as

$$\mathbf{P}_w = \mathbf{R}_{\text{src}}^\top (\mathbf{P}_{\text{src}} - \mathbf{T}_{\text{src}}),$$

by inverting the source extrinsic transformation. Substituting this into the target camera equation

$$\mathbf{P}_{\text{dst}} = \mathbf{R}_{\text{dst}} \mathbf{P}_w + \mathbf{T}_{\text{dst}},$$

we obtain

$$\mathbf{P}_{\text{dst}} = \mathbf{R}_{\text{dst}} \mathbf{R}_{\text{src}}^\top (\mathbf{P}_{\text{src}} - \mathbf{T}_{\text{src}}) + \mathbf{T}_{\text{dst}},$$

which simplifies to

$$\mathbf{P}_{\text{dst}} = \mathbf{R}_{\text{ref}} \cdot \mathbf{P}_{\text{src}} + \mathbf{T}_{\text{ref}},$$

where the relative rotation and translation are given by:

$$\mathbf{R}_{\text{ref}} = \mathbf{R}_{\text{dst}} \cdot \mathbf{R}_{\text{src}}^\top = \mathbf{R}_{\text{dst}} \cdot \mathbf{R}_{\text{src}}^{-1}, \quad \mathbf{T}_{\text{ref}} = \mathbf{T}_{\text{dst}} - \mathbf{R}_{\text{ref}} \cdot \mathbf{T}_{\text{src}}.$$

This transformation allows points to be converted from the source to the target camera frame. However, in our method, we only use $\mathbf{RT}_{\text{ref}}$. We show that the relative transformation expressed in vector form:

$$\mathbf{P}_{\text{dst}} = \mathbf{R}_{\text{ref}} \cdot \mathbf{P}_{\text{src}} + \mathbf{T}_{\text{ref}},$$

where

$$\mathbf{R}_{\text{ref}} = \mathbf{R}_{\text{dst}} \cdot \mathbf{R}_{\text{src}}^\top = \mathbf{R}_{\text{dst}} \cdot \mathbf{R}_{\text{src}}^{-1}, \quad \mathbf{T}_{\text{ref}} = \mathbf{T}_{\text{dst}} - \mathbf{R}_{\text{ref}} \cdot \mathbf{T}_{\text{src}},$$

is equivalent to the transformation in homogeneous coordinates:

$$\mathbf{P}_{\text{dst}} = \mathbf{RT}_{\text{ref}} \cdot \mathbf{P}_{\text{src}}, \quad \mathbf{RT}_{\text{ref}} = \mathbf{RT}_{\text{dst}} \cdot \mathbf{RT}_{\text{src}}^{-1}.$$

We define the extrinsic matrices in homogeneous coordinates as:

$$\mathbf{RT}_{\text{src}} = \begin{bmatrix} \mathbf{R}_{\text{src}} & \mathbf{T}_{\text{src}} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \quad \mathbf{RT}_{\text{dst}} = \begin{bmatrix} \mathbf{R}_{\text{dst}} & \mathbf{T}_{\text{dst}} \\ \mathbf{0}^\top & 1 \end{bmatrix}.$$

The inverse of the source extrinsic matrix is:

$$\mathbf{RT}_{\text{src}}^{-1} = \begin{bmatrix} \mathbf{R}_{\text{src}}^\top & -\mathbf{R}_{\text{src}}^\top \mathbf{T}_{\text{src}} \\ \mathbf{0}^\top & 1 \end{bmatrix}.$$

Multiplying the destination extrinsic matrix by the inverse of the source yields:

$$\begin{aligned} \mathbf{RT}_{\text{ref}} &= \mathbf{RT}_{\text{dst}} \cdot \mathbf{RT}_{\text{src}}^{-1} \\ &= \begin{bmatrix} \mathbf{R}_{\text{dst}} \cdot \mathbf{R}_{\text{src}}^\top & \mathbf{T}_{\text{dst}} - \mathbf{R}_{\text{dst}} \cdot \mathbf{R}_{\text{src}}^\top \cdot \mathbf{T}_{\text{src}} \\ \mathbf{0}^\top & 1 \end{bmatrix}. \end{aligned} \tag{10}$$

This is equivalent to:

$$\mathbf{RT}_{\text{ref}} = \begin{bmatrix} \mathbf{R}_{\text{ref}} & \mathbf{T}_{\text{ref}} \\ \mathbf{0}^\top & 1 \end{bmatrix},$$

which confirms that the relative transformation defined via matrix multiplication in homogeneous coordinates is consistent with the component-wise formulation. so we can get

$$\mathbf{P}_{\text{dst}} = \mathbf{RT}_{\text{ref}} \cdot \mathbf{P}_{\text{src}},$$

B Additional Related Work

Recent Advancements in DiT. The Diffusion Transformer (DiT) series Peebles & Xie (2023) has demonstrated remarkable performance in image generation by replacing the traditional U-Net denoiser with a pure Transformer architecture, achieving superior scalability and modeling capacity. Building upon this, U-DiT Tian et al. (2024) incorporates the hierarchical structure of U-Net by introducing a QKV-based downsampling mechanism, significantly reducing computational costs while maintaining high generation quality. In the video domain, MaskDiT Qi et al. (2025) and DiTCtrl Cai et al. (2024) enhance fine-grained alignment and flexibility in video generation by introducing symmetric binary masks and multi-prompt handling capabilities, respectively.

C Background and Problem Formulation

Transformer-based autoregressive architecture Inspired by the success of sequential modeling in reinforcement learning Chen et al. (2021) and the transformer architecture's inherent suitability for sequence modeling Vaswani et al. (2017), recent methods Rombach et al. (2021); Ren & Wang (2022) propose flattening the input image $\mathbf{I}_0$ and the relative camera $\mathbf{RT}_{ref}$ into tokens, concatenating them in sequence, and feeding them into a Transformer for long-horizon view synthesis. The model then autoregressively predicts the next-view image I_1 from tokens, and proceeds to generate the full sequence $\{I_2, \ldots, I_T\}$ through image-level autoregression, as formulated below:

$$\begin{aligned} p\left(\{I_t\}_{t=1}^T \mid I_0, \{\mathbf{RT}_t\}_{t=1}^T\right) = \prod_{\tau=1}^{T} \prod_{i=1}^{HW} p\left(I_{\tau,i} \mid I_{\tau,<i}, I_{<\tau}, \right. \\ \left. \{\mathbf{RT}_t\}_{t=1}^T\right) \end{aligned} \tag{11}$$

where $\tau \in [1, T]$ indexes the frame number and $i \in [1, HW]$ indexes the token position within a single frame. These two levels of conditioning correspond to **image-level** and **token-level** autoregression. Most methods adopt a GPT-2-like architecture to model the sequential process:

$$I_t \sim p\left(I_t \mid I_1, \mathbf{RT}_2, I_2, \mathbf{RT}_3, \ldots, I_{t-1}, \mathbf{RT}_t\right). \tag{12}$$

This autoregressive modeling approach has limitations. Since the model relies only on previously generated tokens at each step, early prediction errors can propagate and accumulate, leading to a

lack of global consistency. Additionally, the left-to-right generation process struggles to effectively capture spatial structures under large viewpoint changes, limiting the diversity of generated content. In particular, in regions far from the source view, the model is prone to "dead zones" and may generate rigid or unnatural content. More qualitative results are provided in D.

U-Net Diffusion architecture Recent diffusion models Ho et al. (2020); Song & Ermon (2019) have emerged as powerful tools for sampling images from complex conditional distributions, such as text-to-image generation and image restoration Rombach et al. (2022); Kawar et al. (2022); Zhu et al. (2023); Saharia et al. (2022). Their ability to achieve complex sampling with relatively simple training objectives Sohl-Dickstein et al. (2015); Vincent (2011) makes them an excellent candidate for NVS. In the task we explore, the training of the diffusion model requires conditioning on $\mathbf{c} = [\mathbf{I}_{\text{src}}, \mathbf{RT}_{\text{ref}}, \mathbf{K}, \mathbf{3D}_{\text{info}}]$ and matching it with the target image I_{dst}. This is done by adding Gaussian noise to the sample and training a denoising network ϵ_θ (using the U-Net architecture) to remove the noise from the given conditional signal, thus enabling the model to learn the denoising ability based on c:

$$\text{noise}_i \sim \epsilon_\theta(\mathbf{c}), \tag{13}$$

where $i \in [1, T]$ represents the steps of adding and removing noise. During the denoising process, we can then obtain:

$$I_{\text{dst}} \sim p(I_{\text{dst}} \mid \mathbf{c}). \tag{14}$$

Recent work has categorized diffusion models into two main approaches:

w/o $\mathbf{3D}_{\text{info}}$: These methods Yu et al. (2023); Tseng et al. (2023b); Elata et al. (2024) use implicit modeling without explicitly inputting 3D bias information. However, due to the local modeling of convolutions, they often lack strong global semantic consistency constraints, which may result in regions that are inconsistent with scene logic or semantics (i.e., "unknown blocks").

w/ $\mathbf{3D}_{\text{info}}$: These methods inject additional 3D bias information, such as $\mathbf{I}_{\text{warp}}$ or $\mathbf{I}_{\text{depth}}$, derived from $\mathbf{I}_{\text{src}}$ and $\mathbf{RT}_{\text{ref}}$, into the model Seo et al. (2024); Müller et al. (2024). However, this often leads the model to "complete" the target image I_{dst} based on the input 3D bias, sacrificing effective inter-frame modeling capabilities. For example, Seo et al. (2024) struggles with accurately handling small-scale camera movements.

Diffusion Models. Diffusion models synthesize data by learning to invert a fixed noising process. Given a clean sample x_0, the forward process gradually corrupts it with Gaussian noise:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\Big(x_t;\ \sqrt{1-\beta_t}\, x_{t-1},\ \beta_t I\Big), \tag{15}$$

where β_t is the noise schedule. After T steps, x_T approaches a standard Gaussian.

The reverse process is parameterized by a neural network ϵ_θ, which predicts the noise added at each timestep. The mean of the reverse dynamics can be written as

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\, \epsilon_\theta(x_t, t) \right), \tag{16}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

Training is performed by matching the predicted noise to the noise injected during the forward process. Given the reparameterization

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{17}$$

the training objective becomes

$$\mathcal{L} = \mathbb{E}_{x_0, t, \epsilon}\Big[\big\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon,\ t) \big\|_2^2 \Big]. \tag{18}$$

Sampling begins from Gaussian noise x_T and iteratively applies the reverse transition to obtain x_0.

D Additional Qualitative Results

To more intuitively evaluate the performance of our model, we also provide additional generated samples on the MP3D and Re10K test sets, as shown in Fig. 7 and Fig. 6.

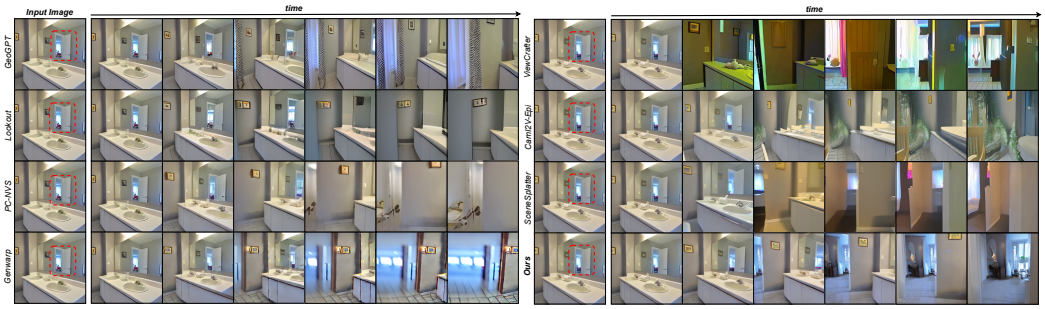

Figure 5: **Qualitative comparison of extrapolated unknown regions.** Continuous novel view synthesis results from different methods given the same input image and camera trajectory. AttenDiffCrafter shows superior visual quality and temporal consistency.



Figure 6: **Qualitative comparison in MP3D test set.** Alternative methods either suffer from local minima or lack temporal continuity and high visual quality. In contrast, our method is capable of generating plausible novel-view sequences that are both temporally and spatially consistent.

E EPIPOLAR VARIANT DESIGN.

When a ray from a point in 3D space is projected onto two different viewpoints, it creates image points that are constrained by epipolar geometry. As shown in the figure, two camera centers O_1 and O_2 define the baseline. The baseline intersects with image planes at points e_1 and e_2, which are called epipoles. The planes containing O_1, O_2, and P are epipolar planes, and their intersections with image planes form epipolar lines. If a world point P projects to point P_1 in i_{src}, then its corresponding point in i_{dst} must lie on the epipolar line e_2P_2. This epipolar constraint significantly reduces the search space for point correspondences.

As shown in Fig 8,given cameras intrinsic matrices K, and relative transformation $R_{\text{ref}}, T_{\text{ref}}$, a world point projects to $P_1(u_1, v_1, 1)$ (homogeneous) in i_{src} and $P_2(u, v, 1)$ (unknown) in i_{dst}. First, we transform to camera coordinates:

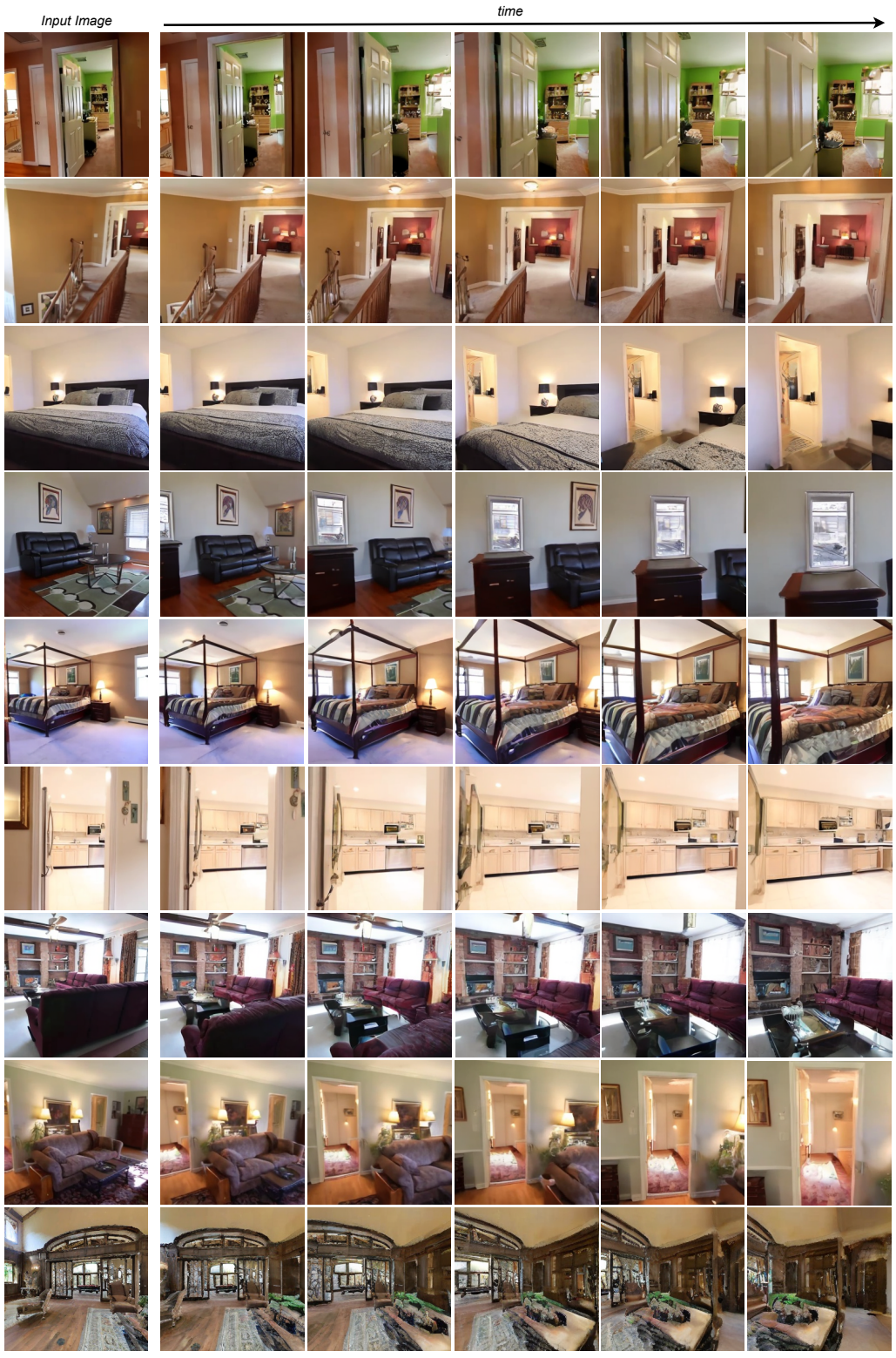

Figure 7: **Additional samples.** We provide the generation results of the first few frames of our generated samples as a demonstration.

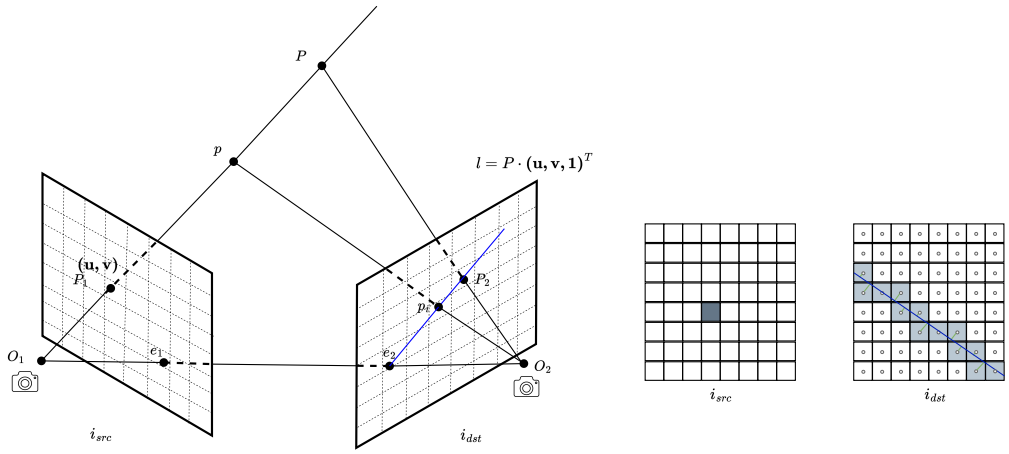

Figure 8: **Epipolar line and mask**. Left: Epipolar constraint of the i_{dst} from one pixel at (u, v) on the i_{src}.Right: Epipolar mask adaptive to the feature in i_{dst}.

$$P_{c1} = \begin{bmatrix} x_{c1} \\ y_{c1} \\ z_{c1} \end{bmatrix} = R_{\text{ref}}(K^{-1} \begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix}) + T_{\text{ref}}$$

$$P_{c2} = \begin{bmatrix} x_{c2} \\ y_{c2} \\ z_{c2} \end{bmatrix} = K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

The normal vector of the epipolar plane O_1O_2P in i_{dst} coordinate system is:

$$\vec{n} = R_{\text{ref}} \times p_{c1} = T_{\text{ref}}^{[T]} p_{c1}$$

where $T_{\text{ref}}^{[T]}$ is the skew-symmetric matrix of T_{ref}. Since p_{c2} lies in this plane, $\vec{p}_{c2} \cdot \vec{n} = 0$.

Substituting and rearranging:

$$[u, v, 1]\, K^{-T} \cdot T_{\text{ref}}^{[T]} R_{\text{ref}} \cdot K^{-1} \begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix} = 0$$

This equation represents the epipolar line for P_2. The matrix $T_{\text{ref}}^{[T]} R_{\text{ref}}$ is known as the Essential Matrix, while $K^{-T} \cdot T_{\text{ref}}^{[T]} R_{\text{ref}} \cdot K^{-1}$ is the Fundamental Matrix.

Epipolar Attention(EA). After mapping each token's feature from i_{src} to 2D coordinates, we can derive the corresponding epipolar line l in i_{dst}. We then calculate the distance from the center of each token's square patch in i_{dst} to line l.

Assuming the epipolar line l in i_{dst} is represented by the equation $au + bv + c = 0$, we compute the distance from each token's center (u_t, v_t) to line l as follows:

First, we obtain the epipolar line coefficients (a, b, c) from the fundamental matrix F and the source point $(u_1, v_1, 1)$:

$$(a, b, c)^T = F \cdot (u_1, v_1, 1)^T$$

where $F = K_2^{-T} \cdot t^{[T]} R \cdot K_1^{-1}$ is the fundamental matrix. Then, for each token center (u_t, v_t) in i_{dst}, we calculate its perpendicular distance to the epipolar line:

$$d = \frac{|au_t + bv_t + c|}{\sqrt{a^2 + b^2}}$$

If this distance satisfies $d < \frac{\sqrt{2}}{2} l_{square}$, where l_{square} is the side length of the token's square patch, we include this token in the attention computation;otherwise,they are masked.This approach yields our Epipolar Attention (EA) mechanism as illustrated in Figure 9, which effectively constrains the attention scope using epipolar geometry.

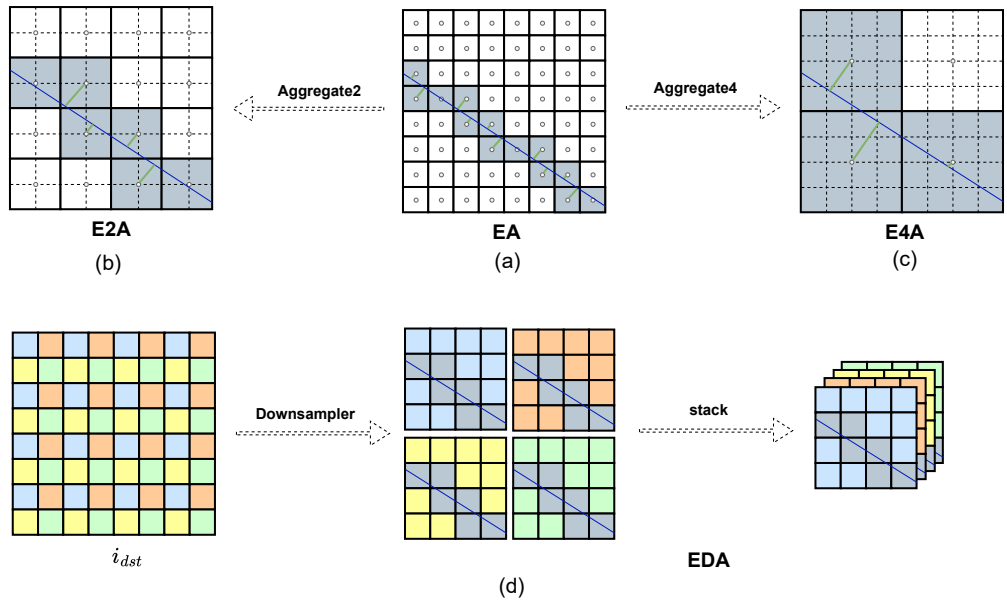

Figure 9: **Epipolar Attention Mechanisms.** Visualization of our proposed attention constraints: (a) basic Epipolar Attention (EA) with individual token evaluation, (b) Epipolar Aggregate-2 Attention (E2A) merging pairs of tokens, (c) Epipolar Aggregate-4 Attention (E4A) combining four adjacent tokens, and (d) DownSampler Epipolar Attention (DEA) which integrates E2A with DownSampler operations to maintain epipolar constraints while preserving model architecture compatibility.

Epipolar Aggragate4 Attention(E4A). To validate our hypothesis that incorporating additional tokens beyond strict epipolar constraints positively impacts our model, we designed an expanded epipolar attention mechanism called Epipolar Aggregate-4 Attention (E4A). This approach extends the basic Epipolar Attention (EA) by merging four adjacent token regions into a single larger region. We then calculate the distance from the center of this enlarged region to the epipolar line to determine whether all tokens within this aggregate region should participate in the attention computation. As illustrated in Figure 9(a,b,c), we demonstrate how the basic EA evolves into E2A (merging two tokens) and further into E4A (merging four tokens). This progressive expansion of the attention field allows us to systematically evaluate the trade-off between strict geometric constraints and beneficial contextual information from neighboring tokens.

DownSampler Epipolar Attention(DEA). However, neither EA nor E4A is compatible with the DownSampler operation, as the downsampling process redistributes points that satisfy epipolar constraints to different feature maps, thereby disrupting the epipolar constraint structure. To address this limitation while maintaining architectural consistency with our model, we developed DownSampler Epipolar Attention (DEA), which simultaneously performs E2A and downsampling operations. As illustrated in Figure 9(d), DEA preserves the epipolar constraint structure as much as possible while accommodating the downsampling operation, ensuring compatibility with our overall model architecture.

F Experiment Details

F.1 Dateset and Metrics

For Re10K, we use video clips from the dataset for training, filtering out shorter sequences, and evaluate the model on its designated test set. As for MP3D, we follow prior works Rombach et al. (2021); Ren & Wang (2022) and use the Habitat simulator Savva et al. (2019) to render both training and testing videos. For both datasets, the videos are resized and center-cropped to a spatial resolution of 256×256.

To measure the discrepancy between the generated images and the ground truth, we use PSNR, SSIM, and LPIPS Zhang et al. (2018). In addition, to better assess the model's capability, we employ FID Heusel et al. (2017) and FVD Unterthiner et al. (2018) to evaluate the visual quality and temporal consistency of the short-term results. These metrics are more sensitive to motion coherence and dynamic fidelity across frames, and better reflect human perception of video smoothness and realism.They also provide a reliable indication of the model's performance in relatively easier regions, such as basic scene synthesis and geometric consistency. We also choose the metric of frame consistency (F-Consistency), which is commonly adopted in video editing Wu et al. (2023), for long-term videos. The vision model used to measure F-Consistency is CLIP ViT-L/14 Radford et al. (2021).

Rotation Distance. Given the ground-truth and estimated rotation matrices R_i^{gt} and R_i^{gen}, the rotation error is defined as

$$\text{Rdist} = \sum_{i=1}^{n} \arccos\left(\frac{\text{tr}\left(R_i^{\text{gen}} R_i^{\text{gt}\top}\right) - 1}{2}\right). \tag{19}$$

Translation Distance. Let T_i^{gt} and T_i^{gen} denote the normalized translation vectors. The translation error is computed as

$$\text{Tdist} = \sum_{i=1}^{n} \left\| T_i^{\text{gt}} - T_i^{\text{gen}} \right\|_2. \tag{20}$$

We adopt Dust3r Wang et al. (2024)(use V100(32G)) instead of COLMAP due to its robustness under feature inconsistencies, which frequently cause pose-estimation failures on generated views.

We implement AttenDiffCrafter using PyTorch platform. The model is optimized with AdamW using a learning rate of 1×10^{-4} and no weight decay. AdamW decouples weight decay from gradient updates, which improves training stability. All our experiments are conducted on 4 NVIDIA RTX 3090 GPUs (24GB each), and the model is trained with a batch size of 32 for approximately 20 to 25 days.

We employ a linear noise scheduler $\{\beta_t\}_{t=1}^{T}$, where $\beta_1 = 1 \times 10^{-4}$ and $\beta_T = 2 \times 10^{-2}$. During training, the total number of diffusion time steps T is set to 1,000. For inference, we use the DDIM Song et al. (2020) sampling strategy with 250 steps to generate high-quality sequences.

G Limitations And Future Work

While the current model is primarily trained on indoor scenes, future work will explore expanding its robustness by training on more diverse datasets, including outdoor settings and synthetic graphics. Additionally, due to consistent camera intrinsics in training data, the generated images reflect these properties, might limiting complex intrinsic variations.These issues represent important directions for future research.

H Impact Statement

Our method enhances the realism and controllability of scene-level novel view synthesis, benefiting applications in virtual reality, film production, and digital heritage preservation. However, like

many powerful generative models, it also poses potential risks of misuse, such as generating fake or misleading visual content. These risks may be amplified by the capability to produce consistent multi-view sequences from a single image. We advocate for responsible deployment of this technology and encourage future research on model transparency, detection, and traceability to mitigate unintended consequences.