# Multi-view Geometry-Aware Diffusion Transformer for Indoor Novel View Synthesis

**Xueyang Kang & Zhengkang Xiang & Zezheng Zhang & Kourosh Khoshelham** *
Faculty of Engineering and Information Technology
The University of Melbourne
Parkville VIC 3010, Australia
alex.kang@kuleuven.be

## Abstract

ecent progress in novel view synthesis of indoor scenes using diffusion models have attracted significant attention, particularly for generating desired poses from a source image. Existing methods can produce plausible views near the input view, but they often fail to extrapolate views far beyond the input perspective. Additionally, achieving a multiview consistent diffusion model typically requires training of computational-heavy 3D priors, limiting scalability to long-range generation. In this paper, we present a transformer-based latent diffusion model that leverages view geometry constraints, including explicitly warped feature maps of the input view as the denoised target view and a conditioning combination of epipolar-weighted source image feature map, Plücker raymap, and camera poses. This approach allows for the extrapolation of consistent novel views, both semantically and geometrically, over long-range trajectories in a single-shot manner. Our model is evaluated on two indoor datasets, ScanNet and RealEstate10K, using a diverse set of metrics for view quality and consistency evaluation. Experimental results demonstrate the superiority of our approach over existing models, showcasing its potential for semantically and geometrically consistent novel view synthesis scalably in video generation.

## 1 Introduction

Existing novel view synthesis (NVS) methods, such as NeRF (Mildenhall et al., 2020), approach view synthesis as an interpolation problem between input views. This approach restricts the synthesis to observed regions and necessitates multiple input views to learn an implicit scene representation. Despite existing NVS methods from single input image have demonstrated promising results in near view to the reference image, nonetheless, they struggle to extrapolate consistent views in long range, and usually require a 3D geometry prior as guidance, including depth, NeRF prior or fusion of multiview feature embeddings. For example, SE3DS (Koh et al., 2023) generates new views by using a reprojected point cloud as a guidance. Invisible Stitch (Engstler et al., 2024) iteratively stitches and fuses newly generated views through an on-the-fly depth completion model. For the work using NeRF as 3D prior, Gaudi (Bautista et al., 2022) optimizes a latent representation by decoupling the scene radiance field and camera poses to render consistent views along a specified trajectory. ZeroNVS (Sargent et al., 2024) leverages the NeRF prior to guide diffusion through score distillation sampling with anchoring views. For diffusion models conditioned by multiview embeddings, MVDiffusion (Tang et al., 2023) enhances multiview interactions by integrating correspondence-aware attention layers, generating all views simultaneously. CAT3D (Gao et al., 2024) learns 3D representations conditioned on multiview images and target viewpoints.

- We introduce a novel diffusion transformer constrained by view-to-view geometry via interleaving of self-attention and cross-attention, showing consistent generation in long-range indoor views.

*Xueyang is also with PSI division of the department of Electrical Engineering (ESAT) of KU Leuven, 3000 Leuven, Belgium.

- We propose a novel adaptive fusion noise score (FNS) to dynamically adjust the conditioning effects according to the ratio of warped regions over the feature map.
- To the best of our knowledge, we are the first to apply view-geometry constraints in the latent space of the diffusion model. This accelerates training and improves NVS robustness as compared to the methods that directly leverage geometry constraints in pixel space.

## 2 Related Work

**Diffusion for Image Generation.** The diffusion model, particularly the denoising diffusion model (Ho et al., 2020), has revolutionized image generation in computer vision (Gu et al., 2022; Wang et al., 2022; Chefer et al., 2023; Liu et al., 2023), expanding to domains including image inpainting, while offering superior training stability compared to generative adversarial networks (GANs). The introduction of classifier-free guidance (Ho & Salimans, 2021) benefits conditional diffusion models. The latest CAT3D (Gao et al., 2024) presents a promising approach to novel view synthesis and 3D scene reconstruction using multiview diffusion model, it primarily focuses on scenarios where multiple input views are available and leverages these to create consistent 3D representations.

**Novel View Synthesis from a Single Image.** Conditioning diffusion models on a single view of a complex scene presents challenges that go beyond object-centric view synthesis, such as handling out-of-distribution camera poses, large-scale scenes with occlusions. Despite these challenges, diffusion models have been applied to 3D scene generation (Rockwell et al., 2021; Bahmani et al., 2023; Kim et al., 2023; Höllein et al., 2023; Huang et al., 2023; Koh et al., 2023; Tang et al., 2024). GeoGPT (Rombach et al., 2021), the first transformer-based model for synthesizing image sequences along a trajectory, explored explicit, implicit, and hybrid geometric priors derived from depth. Look Outside the Room (Ren & Wang, 2022) introduced a local constraint on input cameras to progressively generate views along a trajectory, using a concurrent diffusion model for stochastic conditioning in an autoregressive manner. However, these autoregressive models (Rombach et al., 2021; Ren & Wang, 2022) are susceptible to blurring and drift errors. The latest long-term Photometric-consistent NVS (Yu et al., 2023) employs a diffusion model for view-to-view translation, facilitating single-shot view generation through cross-attention between the source and target view streams, respectively. MultiDiff (Müller et al., 2024) integrates monocular depth priors to warp source views as references and video diffusion piror (Chen et al., 2024) to produce multiview consistency results for long-term scene generation with large camera movements. Despite these advancements, handling out-of-distribution camera poses, occlusions, and large-scale scenes remains challenging.

## 3 Method

The single-shot novel view synthesis (NVS) problem using a diffusion model can be formulated as sampling a view image from the conditional distribution $p_\theta$.

$$p_\theta(x_j|x_i, \mathbf{R}_i^j, \mathbf{t}_i^j, \mathbf{K}), \tag{1}$$

where $x_i \in \mathbb{R}^{H\times W\times C}$ is the source input image, $\mathbf{K}$ represents the camera intrinsic parameters, and $(\mathbf{R}_i^j, \mathbf{t}_i^j)$ denotes the relative transformation from the source view to the target view. Here, $C$ stands for the number of channels, while $H$ and $W$ represent the height and width, respectively. Sampling from this conditional distribution introduces ambiguities due to multiple possible realizations. When the target view $x_j$ is significantly distant from the source view $x_i$, the overlap may be minimal, limiting its contribution to the sampling process and complicating the sampling distribution task. Our model, as illustrated in Fig. 1, leverages relative view geometry as guidance to address these challenges. It is designed to manage view ambiguity when there is minimal overlap between the source and target views. The model seamlessly integrates the conditioning semantic features of the source image $x_i$ with the warped feature map of $x_j^{\text{known}}$, derived from the source image using depth and relative camera transformation. For efficient diffusion training, all images are mapped into latent features $z$ using a pre-trained variational autoencoder (VAE) (Kingma & Welling, 2013), preserving the spatial structure of the images. All subsequent geometric transformations are applied to $z$ before converting the feature map into tokens.

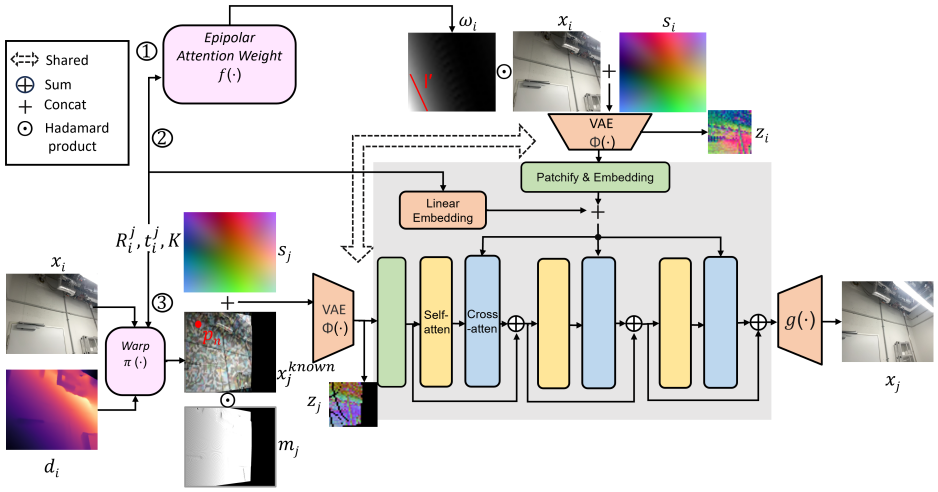


Figure 1: Our single-shot diffusion model synthesizes the target image $x_j$ by taking the warped image $x_j^{\text{known}}$, derived from depth-based mapping $\pi(\cdot)$ of the source image $x_i$, as input, with $x_i$ serving as the condition. The model incorporates interleaving self-attention (orange) and cross-attention (blue) layers to effectively generate using warping signals and conditional features. Only the latent diffusion layers (gray box) are trained, utilizing a pre-trained VAE encoder $\phi(\cdot)$ and decoder $g(\cdot)$. Relative camera transformations $(\mathbf{R}_i^j, \mathbf{t}_i^j)$ and intrinsics $\mathbf{K}$ are employed for (1) computing the epipolar attention weight mask $\omega_i$, and the target view is scaled by the binary mask $m_i$ of warped pixels for epipolar attention calculation, (2) embedding camera parameters, and (3) depth map-based warping. The Plücker raymap embedding ($s_j$ or $s_i$) is concatenated with each view image. The epipolar weight mask $\omega_i^k$ is initialized from pixel $p_n$ using the fundamental matrix and is mapped to the corresponding epipolar line $l'$ for illustration on pixels (virtually on the latent feature map $z_i$ and $z_j$).

### 3.1 Preliminaries

We use denoising diffusion probabilistic models (DDPMs) schedules for diffusion. To leverage the learning efficiency of feature map dimension reduction, we implement all the following work in the latent space. The diffusion model starts with the forward process of adding Gaussian noise iteratively at discretized timesteps $t \in \{1, \ldots, T\}$,

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t I), \tag{2}$$

where $z_t$ is the noising feature map of the generated target view image, and $I$ is an identity matrix. $\beta_t$ is a gain dependent on the forward process. Correspondingly, the reverse denoising parameterized by $\theta$ has a standard Gaussian form:

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)), \tag{3}$$

where $\mu_\theta(z_t, t)$ and $\Sigma_\theta(z_t, t)$ are mean and variance respectively. $z_{t-1}$ is derived from the equation below:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(z_t, t) \right) + \sigma_t u_t, \tag{4}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$, and $u_t \sim \mathcal{N}(0, I)$. $\epsilon_\theta(x_t, t)$ states the score function for estimation of noise to denoise feature map $z_t$ into $z_{t-1}$. The equation depicts the iterative update rule for the latent variable $z_t$ in a diffusion model. The diffusion score function $\epsilon_\theta(z_t, t)$ represents the estimation of the noise that needs to be removed from the current latent variable $z_t$ in order to denoise it and recover the corresponding clean feature map. Specifically, the diffusion score function $\epsilon_\theta(z_t, t)$ is predicted by a neural network that takes the current latent variable $z_t$ and the timestamp $t$ as input, and outputs an estimate of the noise distribution at that timestamp $t-1$. This noise estimate is then used to update the latent variable $z_t$ to obtain the next latent variable $z_{t-1}$ in the diffusion process. The term $\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}$ is the scaling factor applied to the denoised predicting output of the score function $\epsilon_\theta(z_t, t)$. The scaling factor applied to the denoised output of the score function decreases proportionally as the time step increases.

### 3.2 DiT Block

Our model is primarily composed of Diffusion Transformer (DiT) blocks, with each block consisting of three key cascaded modules. The DiT block processes tokenized inputs of feature maps through a sequence of layers, beginning with Layer Normalization, followed by Multi-Head Self-Attention, Multi-Head Cross-Attention, and a Pointwise Feedforward Network. The self-attention mechanism within the DiT block captures intra-sequence dependencies, while the cross-attention module integrates conditional context, making the model conditioned by additional context inputs. Outputs from each layer are modulated by a conditioning mechanism via a Multi-Layer Perceptron (MLP), which adjusts the scale and shift parameters of feature distribution, allowing the model to learn view-to-view transformation effectively. This modular design, with attention and conditioning integrated at every step, enables the DiT block to progressively refine and inpaint the input data.

### 3.3 Epipolar Attention

To ensure view geometry consistency, we implement an epipolar attention weight mask inspired by (He et al., 2020; Tseng et al., 2023). The weight mask multiplicatively scales the source image feature map based on the point-to-epipolar-line distance weights. A binary mask, initialized from the valid warped pixels of the source view, further refines the epipolar mask (represented as $m_i$ in Fig. 1). This ensures that the corresponding epipolar weights are computed only within visible regions of the warped reference view, excluding missing or occluded areas in target view. For brevity, in the following clarification, we omit the subscript index, temporal step $t$ for feature map $z$, and the subscript is thus only used for the view index $i/j$. The cross attention is defined as the Hadamard product (elementwise product) $z_i \cdot z_j$ of source and target feature maps $z_i, z_j \in \mathbb{R}^{H' \times W' \times C'}$.

### 3.4 Feature Map Warp

The warp operation $\pi(\cdot)$ is performed on the downscaled latent feature map resolution of the source image. Feature map warping can be described as mapping pixel locations ($m \in 1, \ldots, W', n \in 1, \ldots, H'$) in the source feature map ($z_i \in \mathbb{R}^{H' \times W' \times 3}$) to locations $(m', n')$ in the target feature map using the corresponding depth value $\tilde{d}_i(m, n)$, which is scaled proportionally to match with the feature map dimension of $H' \times W'$. The homogeneous coordinate relationship for warping can be formulated as:

$$\begin{pmatrix} m' \\ n' \\ 1 \end{pmatrix} \simeq \mathbf{K} \left( \mathbf{R}_i^j \mathbf{K}^{-1} \tilde{d}_i(m, n) \begin{pmatrix} m \\ n \\ 1 \end{pmatrix} + \mathbf{t}_i^j \right). \tag{5}$$

This warping maps the feature pixels from the source view to the target view, allowing some pixels to be re-observed in the target view while others may be out of view. As the depth used here is either measured by an RGB-D sensor, as in ScanNet (Dai et al., 2017), or predicted through a monocular depth model (Bhat et al., 2023), it contains some uncertainties. Therefore, we use an approximate relationship in Eq. (5).

### 3.5 Condition & Feature Embedding

The source image $x_i$ concatenated with Plücker raymap embedding $s_i$ of the corresponding camera view, is encoded via the shared weights of VAE $\phi(\cdot)$ into a feature map $z_i$, which is subsequently concatenated (concat) with the linear embedding of relative camera parameter $\rho(\mathbf{R}_i^j, \mathbf{t}_i^j)$:

$$\psi(\cdot) = \text{concat}(\phi(\text{concat}(\omega_i \odot x_i, s_i)), \rho(\mathbf{R}_i^j, \mathbf{t}_i^j))). \tag{6}$$

### 3.6 Inpainting Sampling and Adaptive Fusion Noise Score

The denoising of the target view feature map is modeled as an inpainting process (Lugmayr et al., 2022), where we incorporate the warped pixels from the source feature map in the reverse diffusion process. This forms a binary mask $m$ representing known regions, while the inverse mask $(1 - m)$

indicates the unknown regions. For brevity, in the following part, we omit the subscript index of view id $i/j$ for feature map $z$, and the subscript is used for the time step $t$ index instead.

$$z_{t-1}^{\text{known}} \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_t} z_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{7a}$$

$$z_{t-1}^{\text{unknown}} \sim \mathcal{N}\left(\mu_\theta(z_t, t), \Sigma_\theta(z_t, t)\right), \tag{7b}$$

$$z_{t-1} = m \odot z_{t-1}^{\text{known}} + (1-m) \odot z_{t-1}^{\text{unknown}}, \tag{7c}$$

$z_{t-1}^{\text{known}}$ is sampled via the known pixels in the warped feature map $m \odot z_0$, while $z_{t-1}^{\text{unknown}}$ is sampled from the diffusion model given the intermediate feature map $z_t$ at time $t$. These unknown and known regions are then combined into the feature map $z_{t-1}$. Lastly, $\odot$ denotes Hadamard product operation. Since the original inpainting (Lugmayr et al., 2022) struggles with harmonizing the generated unknown regions with the known regions, we incorporate an adaptive fusion noise score (FNS) to dynamically balance the effects of the input warping and conditioning:

$$\nabla_{z_t} \log p_\theta(z_t) + \gamma \left(\nabla_{z_t} \log p_\theta(z_t|\psi(\cdot)) - \nabla_{z_t} \log p_\theta(z_t)\right). \tag{8}$$

Where $\nabla_{z_t} \log p_\theta(z_t)$ is the unconditional diffusion gradient, $\nabla_{z_t} \log p_\theta(z_t|\psi(\cdot)$ is the conditional diffusion gradient and $\gamma$ is the FNS scale. The corresponding unconditional diffusion model is implemented by removing all conditioning elements, including the source feature, Plücker embedding and camera poses. To fuse the dynamic warping with FNS, we set $\gamma$ to the ratio of unknown feature pixels count over the whole feature map size $\frac{count(1-m)}{count(z_t)}$. For views with larger warped regions (closer views), a lower FNS scale allows the model to focus on the inpainting sampling from the warped input. Conversely, for views with less warping (farther views), a higher FNS scale is used to strengthen the conditioning effects for consistent view generation.

**Algorithm 1** Diffusion-based View Synthesis with Adaptive Fusion Noise Score (FNS)

**Input**: Source image $x_i$, depth-based warping $\pi(\cdot)$, epipolar attention weight $\omega_i$, FNS scale $\gamma$
**Output**: Synthesized target image $x_j$
Initialize latent variable $z_T \sim \mathcal{N}(0, \mathbf{I})$
**for** $t = T$ **downto** $1$ **do**
    **if** $t > 1$ **then**
        Sample noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$;
    **else**
        $\epsilon = 0$;
    **end if**
    $\psi(\cdot) = \text{cat}(\phi(\text{cat}(\omega_i \odot x_i, s_i)), \rho(\mathbf{R}_i^j, \mathbf{t}_i^j)))$;
    $z_{t-1}^{\text{known}}, z_{t-1}^{\text{unknown}} = \Phi(\pi(x_i))$;
    Compute adaptive fusion noise score:
    $\nabla_{z_t} \log p_\theta(z_t) + \gamma \left(\nabla \log p_\theta(z_t|\psi(\cdot)) - \nabla \log p_\theta(z_t)\right)$;

    $\gamma \leftarrow \frac{\text{count}(1-m)}{\text{count}(z_t)}$; *// Update FNS scale based on the ratio of unknown regions to total regions*
    $z_{t-1}^{\text{unknown}} \leftarrow \epsilon_\theta(z_t, t) + \omega$, where $\omega \sim \mathcal{N}(0, \mathbf{I})$
    $z_{t-1} \leftarrow m \odot z_{t-1}^{\text{known}} + (1-m) \odot z_{t-1}^{\text{unknown}}$
    **if** $t > 1$ **then**;
        $z_t \leftarrow \frac{1}{\sqrt{\alpha_t}} z_{t-1} + \frac{1-\alpha_l}{\sqrt{1-\alpha_t}} \epsilon_\theta(z_{t-1}, t-1)$;
    **end if**
**end for**
**Return** $x_0 = g(z_0)$;

Algorithm 1 depicts the diffusion process for the novel view synthesis via adaptive fusion noise scores (FNS) to dynamically balance warping signals and conditional features, to synthesize a target view image $x_j$ harmoniously. Starting with a latent variable $z_T \sim \mathcal{N}(0, \mathbf{I})$, the algorithm iteratively denoises $z_t$ by conditioning known regions derived from depth-based warping $\pi(\cdot)$, epipolar attention weights $\omega_i$, and a binary mask $m_j$ derived from valid warped pixels.

The FNS function adjusts the conditional gradient $\nabla_{z_t} \log p_\theta(z_t|\psi(\cdot))$ relative to the unconditional gradient $\nabla_{z_t} \log p_\theta(z_t)$, scaled by $\gamma$, which is updated adaptively based on the ratio of unknown to total regions. Unknown regions are synthesized using the noise model $\epsilon_\theta$, while known regions

are directly propagated. The combined latent representation $z_{t-1}$ is progressively refined, and the final target image $x_j$ is decoded using $g(\cdot)$. This approach ensures harmonization between known and unknown regions for high-quality synthesis from near to far view.

## 4 Experiments

**Baselines.** We evaluate our approach on two indoor datasets: RealEstate10K (Zhou et al., 2018) and ScanNet (Dai et al., 2017). For evaluation on RealEstate10K (Zhou et al., 2018), we use various baseline models, including Stable Diffusion 2.0 (Rombach et al., 2022a) (SD-Inpainting), Look Outside the Room (Ren & Wang, 2022) (Look Out), an auto-regressive view diffusion model, Simple and Effective Synthesis Model (SE3DS) (Koh et al., 2023), and Long-term Photometric Consistent Novel View Synthesis Model (PhotoNVS) (Yu et al., 2023).

**Evaluation Metrics.** We employed several common metrics to evaluate the quality and consistency of the generated images, including the Fréchet Inception Distance (FID) (Heusel et al., 2017) , Peak Signal-to-Noise Ratio (PSNR) and the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018). Pixel-wise similarity metric SSIM (Wang et al., 2004) was also employed. However, since these metrics are not sensitive to view geometry consistency, we further incorporated the Thresholded Symmetric Epipolar Distance (TSED) proposed by (Yu et al., 2023) to measure the view geometry consistency between source and target views.

**Implementation Details.** Our diffusion model is based on the Diffusion Transformer (DiT) architecture (Peebles & Xie, 2023), featuring interleaved self-attention and cross-attention mechanisms. We opted for the XL model configuration, which consists of 24 DiT blocks. Each block includes self-attention layers, cross-attention layers, and a final linear feedforward layer. The patch size is set to 2. Regarding the feature map side for diffusion, $H' \times W' \times C'$ is configured as $(32 \times 32 \times 4)$, with input images sized at $256 \times 256 \times 3$, cropped from the center of the raw images. The epipolar attention weight mask thresholds $C_1$ and $C_2$ are set to 0.8 and 2.5, respectively.

### 4.1 Quantitative Comparisons

Table 1 provide the quantitative evaluation results and comparisons with baselines on ScanNet (Dai et al., 2017) (upper part) and RealEstate10K (Zhou et al., 2018) (lower part). The performance of our model dominates across all the metrics consistently in short and long-range view synthesis.

The metric results in Table 1 only evaluate the generated individual image quality against ground truth image, while not reflecting the view geometry consistency. Therefore, we further use the TSED metric to evaluate view consistency. When both median epipolar line distance error and and SIFT feature match meet the threshold, a pairwise image frame match is recorded. The match percentage is reported in Fig. 2. We chose two sequences with the most common trajectory existing in the datasets, namely forward-backward motion and orbital motion involving large rotation.

Table 1: Comparison results on ScanNet (Dai et al., 2017) (upper part) and RealEstate10K (Zhou et al., 2018) (lower part). **Orange** indicates the best and **pink** represents the second best.

| Model | Short Range | | | | Long Range | | | |
|---|---|---|---|---|---|---|---|---|
| | FID ↓ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | FID ↓ | LPIPS ↓ | PSNR ↑ | SSIM ↑ |
| Look Out (Ren & Wang, 2022) | 46.30 | 1.95 | 9.26 | 0.16 | 61.20 | 2.16 | 6.72 | 0.12 |
| SE3DS (Koh et al., 2023) | 29.27 | 0.62 | 10.73 | 0.27 | 65.36 | 2.92 | 5.74 | 0.07 |
| SD-Inpainting (Rombach et al., 2022b) | 30.93 | 0.58 | 11.24 | 0.25 | 54.75 | 1.78 | **9.16** | **0.28** |
| VistaDream (Wang et al., 2024) | **20.74** | **0.35** | **13.49** | **0.51** | **50.63** | **1.59** | 8.14 | 0.16 |
| Ours | **13.93** | **0.25** | **15.86** | **0.62** | **22.15** | **0.57** | **10.50** | **0.50** |
| Look Out (Ren & Wang, 2022) | 16.44 | 0.46 | 16.13 | 0.61 | 19.38 | 0.62 | 11.38 | 0.48 |
| SE3DS (Koh et al., 2023) | 28.35 | 0.86 | 9.73 | 0.32 | 35.25 | 1.19 | 8.50 | 0.18 |
| SD-Inpainting (Rombach et al., 2022b) | 32.85 | 1.34 | 10.15 | 0.28 | 56.13 | 2.21 | 6.37 | 0.12 |
| PhotoNVS (Yu et al., 2023) | 15.56 | **0.43** | **16.51** | **0.61** | **18.91** | **0.56** | **11.96** | **0.58** |
| VistaDream (Wang et al., 2024) | **14.76** | 0.57 | 11.06 | 0.40 | 26.37 | 0.75 | 9.32 | 0.38 |
| Ours | **9.30** | **0.11** | **21.16** | **0.73** | **11.42** | **0.21** | **17.16** | **0.64** |

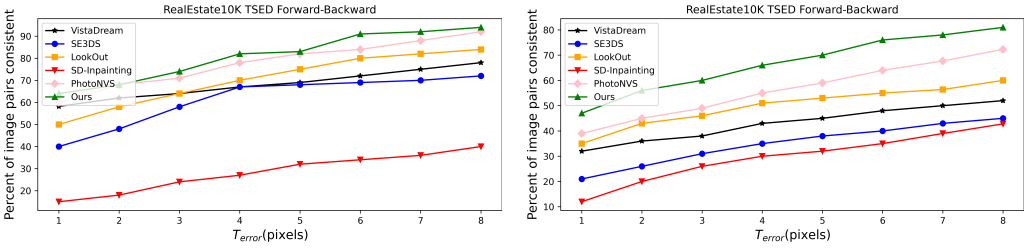


Figure 2: TSED evaluation on RealEstate10K for two dominant motion patterns, forward-backward and orbital trajectory. The TSED plot is percent of consistent image pairs as a function of $\mathbf{T_{error}}$ with $\mathbf{T_{match}}$ set to constant 15.

Figure 3: Baseline comparisons on RealEstate10K (Zhou et al., 2018), where each column is the sampled views of a specific model from near to far, and the source image (highlighted in red box overlaid on the top right image) is close to the first Ground Truth viewpoint at top right.

### 4.2 Qualitative Comparisons

We also provide qualitative comparisons in RealEstate10K (Zhou et al., 2018) using Fig. 3. It is apparent that our model can preserve the view geometry consistency well, while generating new information (as illustrated in the table region of the bottom right GT image) consistent with the semantic context of reference view. VistaDream (Wang et al., 2024), Look Out (Ren & Wang, 2022) and PhotoNVS (Yu et al., 2023) can predict good images at viewpoints close to input view, yet struggle to extrapolate consistent views far away. For example, in the second column, Look Out results degrade to plain walls at far-way viewpoints.

### 4.3 Ablation Study

We report the overall metric results of ablation study in Tab. 2. Here we use the same sequence frames of RealEstate10K (Zhou et al., 2018) without the short and long range splits. The ablation

table results demonstrate that the best overall performance is achieved by the full model leveraging inpainting, conditioning and adaptive FNS (Row 4), which has the lowest FID (10.36), LPIPS (0.16), and median TSED error (24.09) and achieves the highest PSNR (19.16) and SSIM (0.69). These

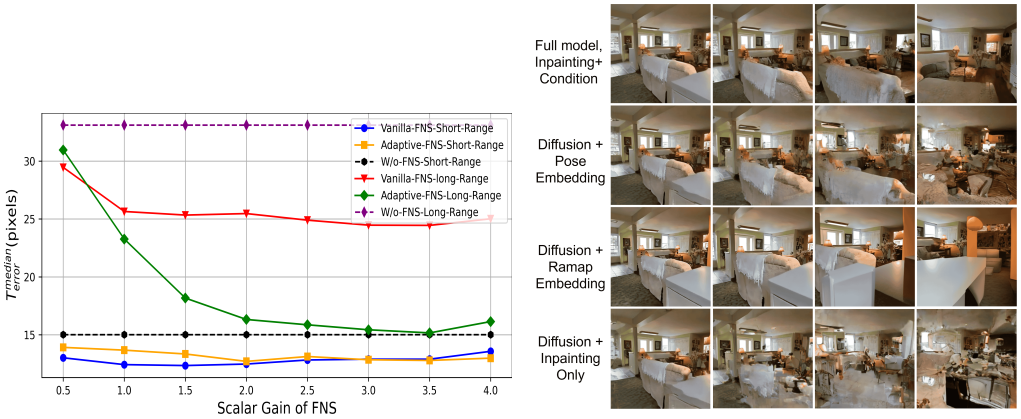


Figure 4: a) Comparison of median translation error (in pixels) between Vanilla Fusion Noise Score (FNS) and Adaptive Fusion Noise Score across short-range and long-range test scenarios. b) Ablation study to evaluate the effectiveness of various modules in pose alignment for generated views.

results show that the geometry-aware conditioning and constraints play a critical role in achieving consistency, smooth transitions between views, even though our method is based on an image diffusion model rather than a video diffusion prior.

Table 2: An ablation study was conducted on our model design, using RealEstate10K (Zhou et al., 2018) as the test dataset. The study involved comparing various input and condition combinations, analyzing relative camera motion geometry constraints.

| | | FID ↓ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | $\mathbf{T_{eror}^{median}}$ ↓ |
|---|---|---|---|---|---|---|
| **1.** | Inpainting only | 15.64 | 0.27 | 15.67 | 0.59 | 48.44 |
| **2.** | Condition only | 11.53 | 0.19 | 18.74 | 0.64 | 29.31 |
| **3.** | Inpainting + Condition | 10.91 | 0.16 | 19.10 | 0.68 | 27.66 |
| **4.** | Inpainting + Condition + FNS | **10.36** | **0.16** | **19.16** | **0.69** | **24.09** |
| **5.** | W/o Epipolar Attention | 10.93 | 0.18 | 19.04 | 0.67 | 28.24 |
| **6.** | W/o Warped Mask for Epi-Atten | 10.62 | 0.18 | 19.10 | 0.67 | 26.85 |
| **7.** | W/o $\mathbf{R}_i^j, \mathbf{t}_i^j$ Embedding | 13.59 | 0.22 | 16.66 | 0.62 | 26.19 |
| **8.** | W/o Plücker Raymap | 13.72 | 0.21 | 16.73 | 0.64 | 25.88 |

The proposed adaptive FNS weight $\gamma$ is a multiplication of scalar gain and the ratio of unknown pixel number in warped feature map size over the whole feature map size. Regarding the scalar gain impact on our adaptive fusion noise score, along with comparison of using vanilla FNS and our adaptive FNS, we provide a plot in Subfig. a of Fig. 4.

(a) Reference (b) Warped Reference (c) Inpainting (d) Condition (e) Cond+Inpaint.

Figure 5: The far-away synthesis view of our model with various structure combinations of inpainting and condition guidance. Reference input view and warped reference view by monocular depth and camera parameters are provided in the first row.

Table 3 highlights minor metric differences between these two sources: GT depth performs better in near views, while ZoeDepth excels in far views. This demonstrates that a reliable monocular depth

prior ensures consistent warping masks for generating coherent views. For the final implementation, we use GT depth for ScanNet, as the evaluation metric differences between ZoeDepth (Bhat et al., 2023) and GT depth are negligible.

Table 3: Comparison results on RealEstate10K (Zhou et al., 2018) for various depth sources.

| Depth | Short Range | | | | Long Range | | | |
|---|---|---|---|---|---|---|---|---|
| | FID ↓ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | FID ↓ | LPIPS ↓ | PSNR ↑ | SSIM ↑ |
| GT Depth | **4.98** | **0.33** | **17.22** | **0.69** | 6.61 | 0.32 | 17.18 | 0.73 |
| Truncated Depth | 5.50 | 0.35 | 17.06 | 0.68 | 10.16 | 0.38 | 16.02 | 0.61 |
| ZoeDepth (Bhat et al., 2023) | 4.99 | 0.33 | 17.21 | 0.69 | **6.57** | **0.32** | **17.19** | **0.73** |

## References


Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Xingguang Yan, Gordon Wetzstein, Leonidas Guibas, and Andrea Tagliasacchi. Cc3d: Layout-conditioned generation of compositional 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7171–7181, 2023.

Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. *Advances in Neural Information Processing Systems*, 35:25102–25116, 2022.

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.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

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*, pp. 5828–5839, 2017.

Paul Engstler, Andrea Vedaldi, Iro Laina, and Christian Rupprecht. Invisible stitch: Generating smooth 3d scenes with depth inpainting. *arXiv preprint arXiv:2404.19758*, 2024.

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

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022.

Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 7779–7788, 2020.

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 and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

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

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7909–7920, 2023.

Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16750–16761, 2023.

Seung Wook Kim, Bradley Brown, Kangxue Yin, Karsten Kreis, Katja Schwarz, Daiqing Li, Robin Rombach, Antonio Torralba, and Sanja Fidler. Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8496–8506, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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, pp. 1169–1178, 2023.

Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 289–299, 2023.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11461–11471, 2022.

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

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.

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.

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.

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*, pp. 14104–14113, 2021.

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 (CVPR)*, pp. 10684–10695, June 2022a.

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, 2022b.

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

Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Shitao Tang, Fuayng Zhang, Jiacheng Chen, Peng Wang, and Furukawa Yasutaka. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint 2307.01097*, 2023.

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, 2023.

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

Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

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.

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.

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.