# MetaView: Monocular Novel View Synthesis with Scale-Aware Implicit Geometry Priors

Yufei Cai[1], Xuesong Niu[2*], Hao Lu[3], Kun Gai[2], Kai Wu[2†], Guosheng Lin[1†]

[1] Nanyang Technological University
[2] Kolors Team, Kuaishou Technology
[3] The Hong Kong University of Science and Technology (Guangzhou)

**Abstract.** Current visual generation models are capable of producing high-quality content, yet they lack a coherent perception of the spatial structure. Existing generative novel view synthesis methods typically introduce explicit geometry priors, which enforce spatial consistency but inherently restrict generalization in large view changes. In contrast, recent interactive generative methods favor implicit scene modeling, offering greater flexibility at the cost of precise camera control and geometry consistency. In this paper, we propose **MetaView**, a diffusion-based monocular novel view synthesis framework that enables rendering under large view changes from a single image. Our key insight is to combine implicit geometry modeling with minimal yet essential explicit 3D cues: we incorporate implicit geometry priors from a feed-forward geometry perception network to regularize structure without imposing restrictive reconstruction pipelines, while leveraging metric depth to anchor the generation to a metric scale. This design allows MetaView to achieve both geometry consistency and precise controllability. Extensive experiments demonstrate that, under challenging monocular large viewpoint changes, MetaView significantly outperforms existing methods and exhibits superior generalization. Our code will be made publicly available.

**Keywords:** Visual Generation · Novel View Synthesis · Camera Control

## 1 Introduction

Benefiting from the proliferation of large-scale visual data, feed-forward paradigms have emerged as an effective and scalable approach to 3D understanding and generation. In 3D perception, recent works [25, 56, 57] show that direct mapping from images to geometric representations offers superior efficiency and robustness compared to traditional optimization-based pipelines. This trend has naturally extended to visual generation, where diffusion-based [15, 27] models reformulate scene reconstruction as a data-driven pixel prediction problem. A pivotal yet challenging task within this domain is monocular novel view synthesis (NVS), which aims to synthesize target views under arbitrary camera poses given a single source input as shown in Fig. 1. While existing methods attempt to address

*Project lead. †Corresponding author.


**Fig. 1: Monocular novel view synthesis results. Upper**: Visual comparisons across different methods. **Lower**: Generalization of our MetaView in diverse scenes.

this by either incorporating explicit 3D representations [38, 66] or learning spatial geometry in a purely implicit manner [14, 49], they often struggle to balance consistency with generalization capability under large viewpoint changes.

As shown in Fig. 2, existing diffusion-based NVS methods generally fall into two categories. One line of research adopts a two-stage pipeline [1, 5, 18, 38, 66]: an explicit reconstruction module first establishes a sparse scene representation, which is then reprojected or warped to the target viewpoint to provide geometric conditions for a diffusion model. Such two-stage pipelines acquire explicit priors through reconstruction, after which the diffusion model mainly performs inpainting rather than spatial reasoning. While these explicit inductive biases ensure local consistency, they are inherently constrained by the quality of sparse reconstruction. Conversely, more recent works [14, 49, 52] attempt to construct visual world models to synthesize novel views in a purely implicit manner by directly conditioning on camera poses. Despite the remarkable progress in interactive generation, their 3D perception (e.g., scene depth) remains ambiguous, leading to a scale drifting issue and spatial inconsistency during camera control.

We argue that enhancing spatial awareness in implicit generation models requires moving beyond this dichotomy. Reliance on explicit reconstruction constrains generalization, while learning 3D structure purely from pixel supervision is underdetermined. To this end, we present MetaView, a diffusion-based framework that achieves precise camera control and high-fidelity synthesis without the burden of explicit 3D reconstruction. Our key insight is to incorporate implicit geometry priors and minimal yet essential 3D inductive biases to regularize the generation process, endowing the model with improved spatial sensitivity and robust monocular novel view synthesis capacity. Specifically, we extract hierarchical features from a feed-forward geometry perception network (e.g., DepthAny-

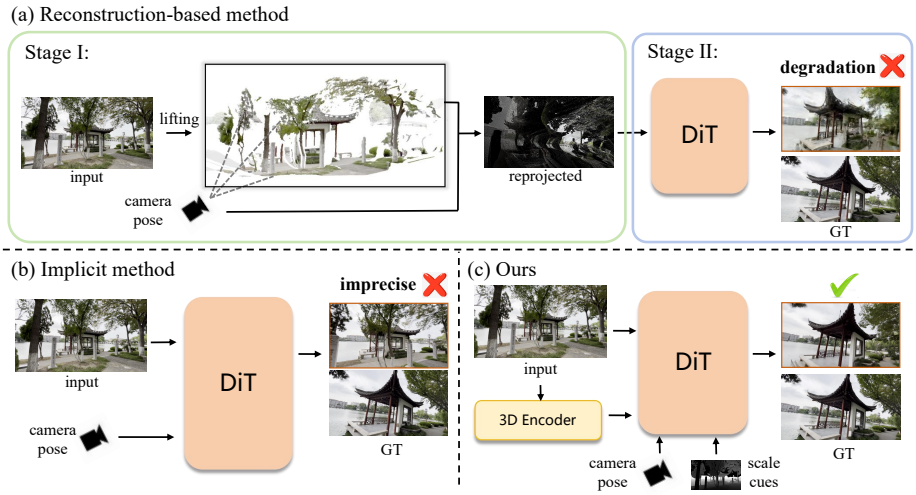


**Fig. 2: Comparison of diffusion-based methods.** Heavy reliance on explicit 3D inductive biases leads to severe visual degradation, while fully implicit methods fail to achieve effective camera control. Our method effectively addresses both issues.

thing3 [25]) as implicit priors to regularize generated scene structure with relative geometry relations. Following PRoPE [23], we encode camera intrinsics and extrinsics into a modified Rotary Positional Encoding (RoPE) [47] to achieve camera control. Simultaneously, to resolve scale ambiguity, we also assign the metric scale cues of the input image into the modified RoPE by allocating an extra subspace for the $z$-axis. Built upon a pretrained multi-modal diffusion transformer (MM-DiT) [12, 61], MetaView incorporates these geometric signals via non-invasive parallel attention layers, preserving the rich semantic knowledge of the base model while injecting structured 3D awareness.

To facilitate a faithful evaluation of NVS under extreme variations, we also introduce a new metric, Dense Matching Distance (DMD), which provides a more accurate assessment of spatial alignment than low-level metrics. We conduct extensive experiments on several datasets, including DL3DV [26], RealEstate10K [72], and Sekai-Real-Walking-HQ [24]. Experimental results demonstrate that, MetaView significantly outperforms both reconstruction-based and purely implicit methods. In particular, our method exhibits superior generalization and more faithful geometric consistency under challenging large viewpoint changes.

Our contributions can be summarized as follows:

- We present MetaView, a robust monocular novel view synthesis method capable of handling large viewpoint changes without explicit reconstruction.
- We introduce implicit geometry priors to facilitate relative geometric consistency in complex scenes. Furthermore, we identify and mitigate the scale drifting issue inherent in implicit generative models by injecting scale cues.
- Extensive experiments demonstrate that MetaView consistently outperforms existing methods and exhibits strong generalization across diverse scenarios.

## 2 Related Works

### 2.1 Text-to-Image Diffusion Models

Diffusion model [15] is a parameterized neural network that learns to approximate complex high-dimensional distributions through an iterative denoising process, and has therefore become widely adopted in visual generation tasks. Latent diffusion model (LDM) [39] performs denoising in a lower-dimensional latent space and introduces conditions through cross-attention mechanisms, achieving a favorable balance between generation efficiency and visual fidelity. Benefiting from large-scale data collection, pretrained text-to-image (T2I) diffusion models have demonstrated remarkable image generation capabilities. Initially, large-scale T2I models [34, 36, 37, 39, 41] adopt U-Net [40] architectures as the denoising backbone. However, as model scales increased, Transformer-based DiT [35] emerged as a more scalable alternative, offering improved generalization and flexibility. Meanwhile, the flow-matching paradigm [13, 27] gained traction due to its streamlined theoretical formulation and stable training dynamics, and has been widely adopted in modern pretrained models. Recent T2I models [4, 12, 21, 61] typically employ a dual-stream Multi-Modal DiT (MM-DiT) [12] architecture that separately models text and image tokens while enabling effective cross-modal interaction. This design substantially improves instruction-following capability and scalability. Our method builds upon Qwen-Image-Edit [61], a variant T2I model tailored for reference image editing. It augments the image stream by concatenating reference image tokens during forward propagation, allowing the model to perform generation conditioned on both text and visual inputs.

### 2.2 Feed-forward Geometry Estimation

Traditional geometry estimation systems [43,44,62,64] decompose 3D reconstruction into handcrafted feature extraction and joint optimization pipelines. Despite the good performance in general scenes, they often suffer from limited robustness and scalability. The emergence of DUSt3R [57] shifts this paradigm by directly performing implicit modeling using a transformer, predicting camera poses and depth maps in a feed-forward manner. Subsequent works [3,51,63,68] extend this framework to handle a larger number of inputs, enabling geometry estimation over broader scenes. VGGT [56] and its variants [10,30,58] further advance feed-forward geometry estimation by leveraging multi-task supervision and large-scale pretraining, substantially pushing the performance to a new level. More recently, DepthAnything3 [25] adopts a minimal yet effective single-transformer architecture, achieving higher efficiency and improved estimation accuracy in multiple downstream tasks. Therefore, we employ the SOTA method DepthAnything3 as our implicit geometry prior network.

### 2.3 Novel View Synthesis

Novel View Synthesis (NVS) [9,22,46] is a long-standing problem in computer vision and graphics. The emergence of Neural Radiance Fields (NeRF) [31] marked

a breakthrough in NVS. NeRF parameterizes 3D representations via implicit neural networks and synthesizes images via volume rendering. Subsequent research [2, 7, 33, 48, 60, 65] focused on improving rendering fidelity and accelerating optimization. 3D Gaussian Splatting (3DGS) [20] improves reconstruction efficiency and rendering quality by representing scenes as anisotropic Gaussian primitives. Later works [8,50,65,67] attempt to loosen requirements on the input views and incorporate neural decoders to enhance generalization.

More recently, diffusion-based generative models have emerged as an alternative framework for NVS [6,28]. ZeroNVS [42] leverages diffusion priors to perform 3D distillation. However, their generalization and geometric consistency remain limited. Subsquent works [1, 45] approache NVS through 2D coordinate warping, but lacks 3D awareness. Other works [5, 18, 38, 66] introduce explicit 3D caches to guide generation via conditioned inpainting. While these methods improve camera control and cross-view consistency, their reliance on explicit reconstruction restricts generalization under sparse-view cases. A different line of work [19, 53, 54, 71] explores large-scale implicit models for scene generation. Recent works [14,49,52] pursue interactive world generation through direct pixel-level regression. Although these methods employ camera pose as a control signal, their spatial geometry awareness remains ambiguous, resulting in imprecise viewpoint control. In this work, we provide the model with spatial awareness by injecting essential 3D cues and implicit geometric priors in a non-invasive manner, striking a balance between controllability and generalization.

## 3 Method

Given a single source image $X^{src}$ and a target camera pose $T$ defined relative to this image, our objective is to synthesize a novel view $X^{gen}$ of the scene $X^{src}$ under the viewpoint specified by $T$. In this section, we first present an overview of the DiT model used in our method (Sec. 3.1). Next, we introduce our approach for injecting explicit 3D cues via RoPE (Sec. 3.2) and explain the extraction and integration method of implicit geometric priors (Sec. 3.3). Finally, we describe the adaptation strategy for the pretrained generation model through parallel attention layers to incorporate these geometric signals (Sec. 3.4).

### 3.1 Preliminary

We adopt pretrained Qwen-Image-Edit [61] as our base model. It is a T2I model built upon the MM-DiT [12] architecture and trained under the flow matching paradigm [13,27], supporting both text and image conditions. The DiT network $v_\theta$ is composed of multiple stacked Multi-Modal Attention (MMA) blocks. Given a reference image and a text prompt, both modalities are first patchified into tokens $X = [X^{src} : X^{gen}] \in \mathbb{R}^{N\times\ d}$ and $C \in \mathbb{R}^{M\times\ d}$, respectively. These tokens are processed in a dual-stream manner, where modality-specific tokens are initially propagated through separate streams. In each MMA block, image and text tokens are projected into modality-specific $Q$, $K$, $V$, which are then concatenated

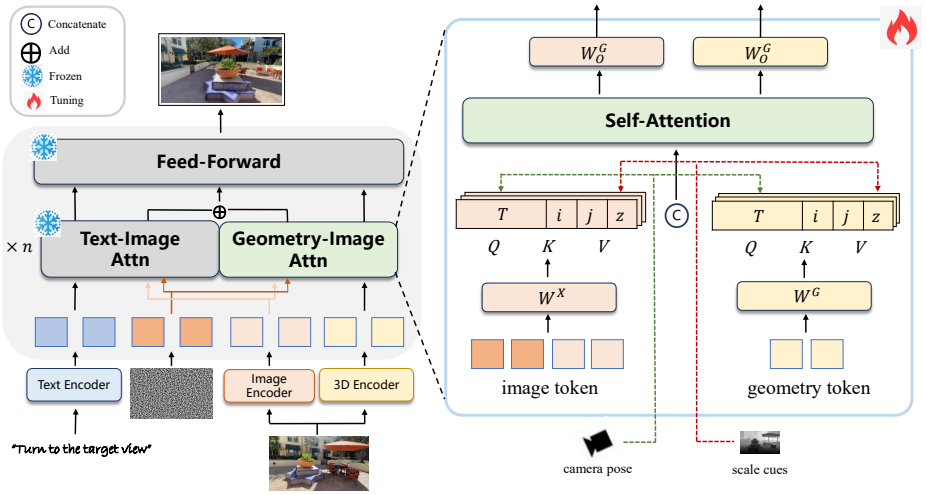


**Fig. 3: Overview of MetaView. Left**: Our method is built upon the MM-DiT architecture, which integrates the implicit geometry priors by adding a split stream in the DiT. **Right**: We fuse geometry priors through self-attention mechanism, while injecting scale cues via a modified spatial-aware RoPE to address the scale drifting issue.

along the sequence dimension and processed via self-attention mechanism [55], enabling cross-modal interaction and information fusion:

$$\text{Attention}([X;C]) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{1}$$

To preserve position relations, tokens are equipped with Rotary Positional Encoding (RoPE) [47]. For text tokens, 1D RoPE is applied to encode relative positions along the sequence. Both source image $X^{src}$ and generated image tokens $X^{gen}$ are assigned rotary transformations by $(k, i, j)$, where $k$ distinguishes different images, and $(i, j)$ denote the token's coordinates on the 2D grid.

### 3.2 3D Cues Injection

The input condition is defined as the set $\{X^{src}, K, T\}$, where $X^{src} \in \mathbb{R}^{H\times W\times 3}$ denotes the input image, $K \in \mathbb{R}^{3\times 3}$ is the corresponding camera intrinsic matrix, and $T = (R, t) \in SE(3)$ represents the relative extrinsic of the target viewpoint. The 6-DoF parameterization of $T$ encodes the camera position and orientation in the world coordinate. Given a 3D point $X_{world} \in \mathbb{R}^4$ in homogeneous world coordinates, its projection onto the image frustum can be formulated as:

$$\tilde{P} = \left[K\ \mathbf{0}^{3\times 1}\right] T, \quad \tilde{X} = \tilde{P} X_{\text{world}} \in \mathbb{R}^3 \tag{2}$$

where $\tilde{X}$ is the projected homogeneous image coordinate. Accurate projection via this formulation requires that $X_{world}$ and $T$ share a consistent metric scale. If the

underlying scale of $X_{world}$ is misaligned with $T$, projection becomes ambiguous, leading to the **scale drifting** issue commonly observed in SLAM systems.

Reconstruction-based methods [1, 18, 38, 66] typically estimate camera poses and depth jointly, ensuring that geometric quantities are aligned within a consistent metric scale. This alignment guarantees geometric consistency during view transformation. In contrast, diffusion-based implicit methods [49, 52] lack an explicit coupling between the provided camera pose and the scene scale internally captured by the model. Since generation models do not explicitly reconstruct 3D structure, the implicit geometry is scale-ambiguous, often resulting in unstable viewpoint control under pose variation.

To mitigate this issue, we introduce the depth $z$ with metric scale and incorporate aligned camera poses during training to eliminate scale ambiguity. Specifically, to enable camera control, we follow PRoPE [23] to encode the intrinsic and extrinsic $K$ and $T$ into RoPE, allowing the model to capture frustum-aware geometric relationships. We adopt the off-the-shelf DepthAnything3-Metric [25] to estimate metric depth $z$ from the input image, explicitly anchoring the generation process to consistent spatial scale cues. In implementation, we allocate an extra subspace within the RoPE to represent the $z$-axis component and assign the extracted $z$ to this dimension. The overall process can be formally expressed as follows:

$$D_t^{\mathrm{RoPE}} = \begin{bmatrix} D_t^{\mathrm{Proj}} & \mathbf{0} \\ \mathbf{0} & D_t^{\mathrm{Grid}} \end{bmatrix}; \quad D_t^{\mathrm{Proj}} = \mathbf{I}_{d/8} \otimes \tilde{P} \in \mathbb{R}^{\frac{d}{2} \times \frac{d}{2}}, \tag{3}$$

$$D_t^{\mathrm{Grid}} = \begin{bmatrix} \mathrm{RoPE}_{d/6}(i_t) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{RoPE}_{d/6}(j_t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathrm{RoPE}_{d/6}(z_t) \end{bmatrix} \in \mathbb{R}^{\frac{d}{2} \times \frac{d}{2}}. \tag{4}$$

where $\otimes$ denotes Kronecker product and $\mathbf{I}$ is identity matrix. $\mathrm{RoPE}_{d/6}(\cdot)$ constructs rotary embeddings in $\frac{d}{6}$ dimensions for $(i, j, z)$ coordinates of token $t$. We follow GTA's [32] formulation to transform $Q$, $K$ and $V$ in the attention:

$$\mathbf{D}^{\mathrm{RoPE}} \odot \mathrm{Attn}\left((\mathbf{D}^{\mathrm{RoPE}})^{\top} \odot Q, (\mathbf{D}^{\mathrm{RoPE}})^{-1} \odot K, (\mathbf{D}^{\mathrm{RoPE}})^{-1} \odot V\right) \tag{5}$$

where $\odot$ denotes matrix product.

Noting that we do not perform handcrafted reconstruction operations such as depth lifting, thereby reducing the extra explicit 3D inductive biases. Unlike methods such as PE-Field [1], we do not rely on fine-grained depth to model relative geometry relationships. Instead, we downsample $z$ to $z'$ to match the spatial resolution of latent tokens in the attention layers and inject sparse cues $z'$ through PE. This design provides an explicit measure of spatial scale, aligning the model's internal spatial perception with the camera geometry via RoPE.

### 3.3 Implicit Relative Geometry Priors

Relative geometric relationships are crucial for novel view synthesis. Although diffusion models can implicitly capture certain relational geometry through pixel-level prediction, object distortions and pose deviations frequently arise in dense

and structurally complex scenes. To mitigate these issues while avoiding excessive explicit 3D inductive biases that may hinder generalization, we incorporate fine-grained implicit relative geometry priors to guide the generation process. Rather than explicit reconstruction, our goal is to provide relative geometry cues that regularize spatial reasoning within the implicit backbone.

Feed-forward 3D perception models [25, 56, 57] trained with multi-task objectives naturally capture rich geometry of scene structure. In addition, the transformer-based architecture produce hierarchical intermediate features that encode geometry in a tokenized form, making it suitable for integration into DiT-based models.

Motivated by this property, we leverage intermediate features from the feed-forward geometry model DepthAnything3 [25] as implicit geometry priors. Specifically, given the source image $X^{src}$, we extract intermediate features from selected layers that are originally fed into the decoder head for 3D prediction. These multi-level features are concatenated along the channel and projected to match the dimension $d$ with image tokens via a linear projection layer $W_G$:

$$G = W_G([\{f^{\phi}\}_{\phi \in decode}]) \in \mathbb{R}^{L \times d} \tag{6}$$

Thus the geometry-awared $G$ are treated as geometry tokens, which are projected to $Q$, $K$, $V$ and then concatenated with image tokens $X = [X^{src}; X^{gen}]$ in self-attention:

$$\text{Attention}([X; G]) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right)V, \tag{7}$$

To ensure consistency, tokens are all encoded under the same RoPE formulation introduced in Sec. 3.2. Similar to source image tokens $X^{src}$, $G$ is associated with the camera pose shifted to the origin, denoted as $T^o = \mathbf{I}$, and is also assigned with the downsampled sparse depth $z'$. In contrast, the generated image tokens $X^{gen}$ corresponding to the target view are bound to the relative target camera pose $T$, while their $z$-axis components are set to 0 to avoid directly constraining the generated scene:

$$\text{PE}(X^{src}) = \text{PE}(G) = \mathbf{D}^{\text{RoPE}}(T^o, z'); \quad \text{PE}(X^{gen}) = \mathbf{D}^{\text{RoPE}}(T, \mathbf{0}) \tag{8}$$

Under this formulation, $G$ inherently encodes relative geometric priors and are additionally anchored with absolute scale through $z'$. The generated image tokens $X^{gen}$, conditioned on the relative transformation $T$, aggregating geometry-aware guidance by attending to tokens defined under $T^o$. This interaction enables geometry-aware generation while preserving the implicit nature of the overall framework.

### 3.4 Geometry-Guided Parallel Attention Adaptation

Our method is built upon a pretrained MM-DiT [12,61] backbone. To maximally preserve its rich semantic knowledge, we freeze all pretrained parameters, thereby preventing catastrophic forgetting. Instead, we introduce parallel attention layers to inject both implicit geometry priors and scale cues.

Specifically, we construct an additional geometry stream for the introduced geometry tokens $G$, which feed forward alongside the original image and text streams across DiT blocks. To maintain compatibility with the original MM-DiT architecture, we do not alter the existing image–text attention. Instead, within each block, we introduce a parallel image-geometry attention layer that models interactions between $X$ and $G$. The resulting features are added to fuse with the original attention outputs, enabling structured 3D cue integration:

$$Out = \text{Attention}([X; C]) + \text{Attention}([X; G]) \tag{9}$$

For the image–geometry attention layer, we introduce a dedicated set of projection parameters $W_Q^X, W_K^X, W_V^X, W_O^X$ and $W_Q^G, W_K^G, W_V^G, W_O^G$ which project the image tokens $X$ and geometry tokens $G$ into their respective query, key, value embeddings and output projection. To ensure parameter efficiency, $G$ and $X$ share the same fixed feed-forward network (FFN) within each transformer block. In addition, a linear projection layer $W_G$ is applied before feeding into the network to map geometry tokens to the same channel dimension $d$ as image tokens. During training, only these newly introduced parameters are optimized, while all pretrained backbone weights remain frozen. The model is trained under the flow matching [27] objective:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,x_0,x_1,C,G}\Big\| v_\theta\big(x_t; C; G\big) - (x_1 - x_0)\Big\|_2^2 \tag{10}$$

## 4 Experiments

### 4.1 Settings

**Datasets**. We conduct experiments on DL3DV [26], RealEstate10K [72], and Sekai-Real-Walking–HQ [24] datasets. We collect only the RGB data from these datasets and estimate camera poses using VIPE [17]. During pose estimation, VIPE adopts DepthAnything3–Metric [25] as scale priors to align poses with metric depth. We further filter out samples with erroneous estimates and low-quality scenes. We evaluate the methods on the three aforementioned datasets, ensuring that test scenes do not overlap with the training set. To assess performance under varying degrees of viewpoint change, we further partition the DL3DV test set into three difficulty levels—easy, medium, and hard—based on the view overlap ratio between the source and target images. Specifically, the easy subset contains pairs with more than 80% overlap, the medium subset includes pairs with 50%–80% overlap, and the hard subset consists of pairs with 30%–50% overlap. For the RealEstate and Sekai test sets, the view overlap ratio is controlled within the range of 30%–80% to ensure diverse yet meaningful viewpoint variations. Details of data preparation are provided in the *Suppl.*

**Implementation details.** We adopt Qwen-Image-Edit [61] as the base model and use DepthAnything3-Giant [25] to extract geometry priors. All experiments are conducted on 8 NVIDIA H800 GPUs with a total batch size of 32. We employ

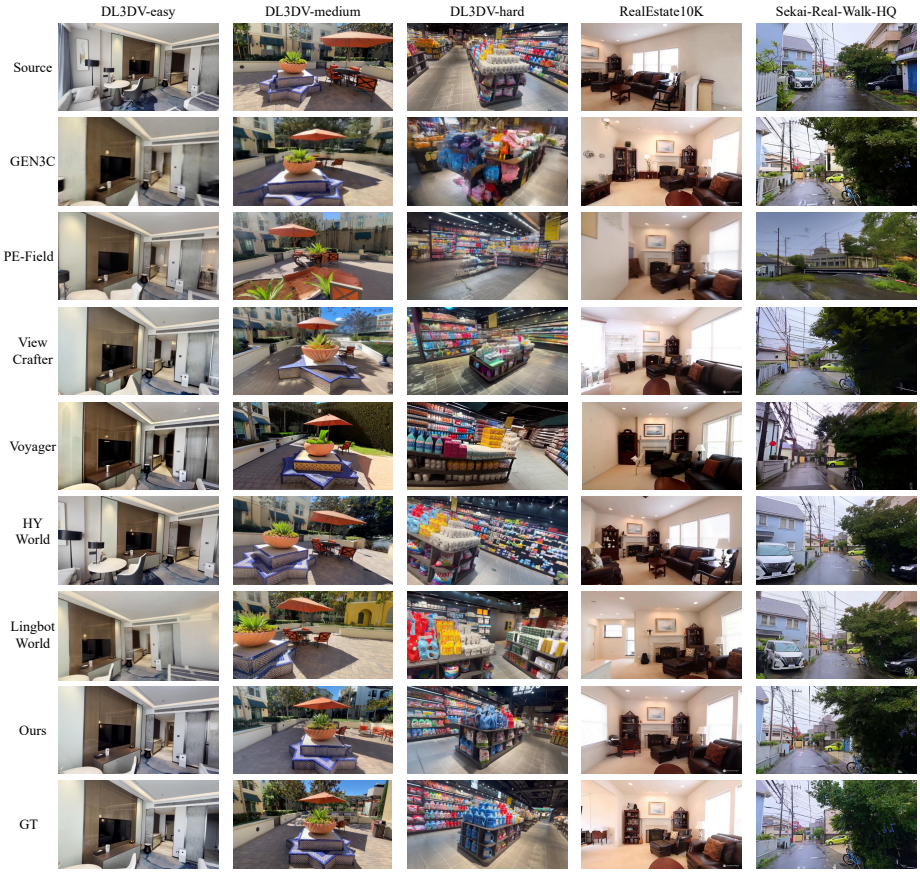


**Fig. 4: Visual comparisons.** Our method enables more precise camera control and significantly outperforms baseline methods under large viewpoint changes.

the AdamW [29] optimizer with a learning rate of 5e-5. The text prompts are fixed to "Turn to the target view". For inference, we use 40 sampling steps with a classifier-free guidance [16] scale of 4.

**Baselines**. We compare our method with both reconstruction-based method: ViewCrafter [66], Gen3C [38], PE-Field [1], Voyager [18], and implicit method: HY-World-1.5 [49], Lingbot-World [52]. All baselines are reproduced using the official open-source implementations with default configurations.

**Metric** We evaluate the methods with low-level metrics PSNR, SSIM [59] and perceptual metric LPIPS [69]. However, the above NVS metrics primarily evaluate performance under narrow viewpoint changes, where novel views are generated via interpolation. When the viewpoint variation becomes large and requires extrapolative generalization, these metrics fail to adequately reflect the capability. Inspired by WorldScore [11], we propose a new metric for NVS under arbitrary viewpoint variations, termed Dense Matching Distance (DMD). It

**Table 1: Quantitative comparisons across difficulty on DL3DV.**

| Method | DL3DV-Easy [26] | | | | DL3DV-Medium [26] | | | | DL3DV-Hard [26] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | DMD↓ | PSNR↑ | SSIM↑ | LPIPS↓ | DMD↓ | PSNR↑ | SSIM↑ | LPIPS↓ | DMD↓ |
| ViewCrafter [66] | 15.45 | 0.4454 | 0.2037 | 8.84 | 13.19 | 0.3679 | 0.2733 | 24.35 | 11.68 | 0.3449 | 0.3223 | 48.83 |
| Gen3C [38] | 15.32 | 0.4339 | 0.1964 | 6.52 | 14.14 | 0.4092 | 0.2556 | 12.43 | 11.87 | 0.3424 | 0.3089 | 41.07 |
| Voyager [18] | 15.19 | 0.4171 | 0.2530 | 12.21 | 13.45 | 0.3610 | 0.2886 | 23.89 | 11.24 | 0.3302 | 0.3462 | 54.26 |
| PE-Field [1] | 15.01 | 0.3964 | 0.2129 | 8.91 | 13.18 | 0.3435 | 0.2749 | 15.41 | 11.97 | 0.3422 | 0.3278 | 41.46 |
| HY-World-1.5 [49] | 12.27 | 0.3323 | 0.2906 | 14.58 | 11.34 | 0.2948 | 0.3304 | 18.50 | 10.71 | 0.3043 | 0.3520 | 34.45 |
| Lingbot-World [52] | 11.82 | 0.3191 | 0.3073 | 29.20 | 11.99 | 0.2912 | 0.3141 | 21.37 | 11.13 | 0.3319 | 0.3447 | 55.39 |
| Ours | **17.94** | **0.5542** | **0.1397** | **2.56** | **15.05** | **0.4225** | **0.2132** | **7.57** | **12.54** | **0.3802** | **0.2881** | **20.74** |

**Table 2: Quantitative comparisons on different datasets.**

| Method | DL3DV [26] | | | | RealEstate10K [72] | | | | Sekai-Real-Walk-HQ [24] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | DMD↓ | PSNR↑ | SSIM↑ | LPIPS↓ | DMD↓ | PSNR↑ | SSIM↑ | LPIPS↓ | DMD↓ |
| ViewCrafter [66] | 13.44 | 0.3861 | 0.2664 | 27.34 | 13.16 | 0.4994 | 0.2597 | 10.92 | 15.73 | 0.4388 | 0.3068 | 17.93 |
| Gen3C [38] | 13.78 | 0.3952 | 0.2536 | 20.01 | 14.07 | 0.5162 | 0.2306 | 9.01 | 16.56 | 0.4659 | 0.2568 | 8.72 |
| Voyager [18] | 13.29 | 0.3694 | 0.2959 | 30.12 | 14.55 | 0.5170 | 0.2519 | 12.95 | 16.24 | 0.4402 | 0.3205 | 19.07 |
| PE-Field [1] | 13.39 | 0.3607 | 0.2719 | 21.93 | 14.51 | 0.5191 | 0.2354 | 14.81 | 15.74 | 0.4418 | 0.2900 | 19.35 |
| HY-World-1.5 [49] | 11.44 | 0.3105 | 0.3243 | 22.51 | 12.34 | 0.4466 | 0.3040 | 13.91 | 14.62 | 0.4088 | 0.3240 | 15.04 |
| Lingbot-World [52] | 11.65 | 0.3141 | 0.3220 | 35.32 | 12.26 | 0.4554 | 0.3026 | 24.51 | 14.52 | 0.3973 | 0.3258 | 18.71 |
| Ours | **15.18** | **0.4456** | **0.2137** | **10.29** | **15.27** | **0.5705** | **0.1980** | **6.44** | **17.72** | **0.5047** | **0.2333** | **6.69** |

measures the average optical flow displacement between matched points across two views. Specifically, we employ UFM [70] to compute the co-visible regions and establish dense correspondences between the two views. Let $\mathrm{UFM}(A; B)$ denote the set of points in $A$ matched to $B$ and their corresponding optical flow displacements $dist(p)$ respect to $B$ computed via UFM, we respectively obtain the matched point sets for the source and generated images relative to the ground truth $P_{src} = \mathrm{UFM}(X^{src}; X^{GT})$ and $P_{gen} = \mathrm{UFM}(X^{gen}; X^{GT})$. For each $p$ in $P_{src}$, if it also exists in $P_{gen}$, its optical flow distance is contributed to the metric; otherwise, a maximum distance penalty $\sigma$ is imposed:

$$\mathrm{DMD}(X^{src}; X^{gen}; X^{GT}) = \frac{1}{|P_{\mathrm{src}}|} \sum_{p \in P_{src}} \begin{cases} \|\mathrm{dist}(p)\|_2^2, & p \in P_{gen} \\ \sigma, & p \notin P_{gen} \end{cases} \tag{11}$$

The reported metric values are normalized to the range of $0 \sim 100$. More analysis about the metric is provided in the *Suppl.*

### 4.2 Qualitative Evaluation

As shown in Fig. 4, MetaView demonstrates effective spatial awareness, producing coherent results under arbitrary viewpoint changes (Cols. 2, 3), while achieving precise control over camera pose and scene scale (Cols. 1, 5). In contrast, fully implicit methods [49,52] suffer from scale drifting and inaccurate viewpoint control due to scale ambiguity. Reconstruction-based methods [1,18,38,66] perform well under narrow viewpoint changes but exhibit blurriness and distortions (Col. 1) due to sparse reconstruction limits. Under large viewpoint variations, these

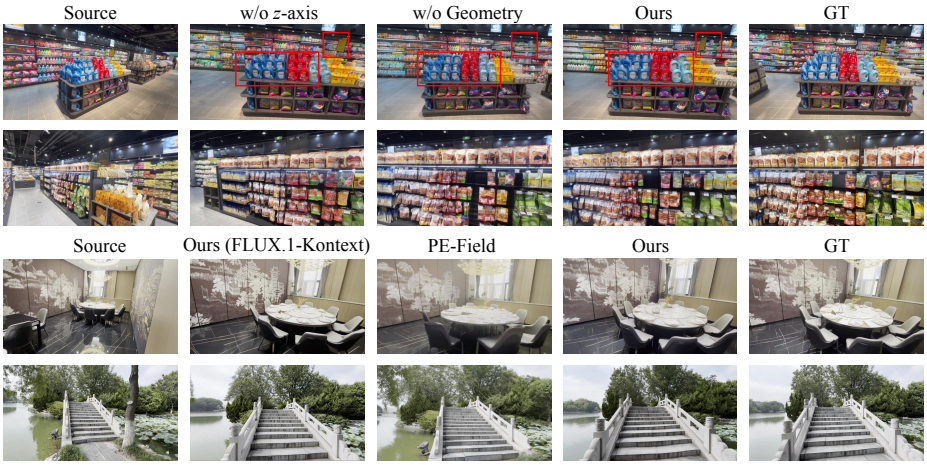


**Fig. 5: Visualization of the ablation.** Scale cues provide the model with accurate spatial awareness, while geometry priors facilitate fine-grained relative consistency.

explicit geometric conditions fail, leading to severe synthesis degradation. Therefore, MetaView outperforms existing methods, demonstrating superior accuracy and generalization. Please refer to the *Suppl* for more results.

### 4.3 Quantitative Evaluation

Tab. 1 presents the quantitative results on the DL3DV dataset across different difficulties. Reconstruction-based methods [1,18,38,66] perform well on the easy subset but degrade significantly as difficulty increases. Implicit methods [49, 52] perform poorly across all subsets due to imprecise viewpoint control, showing limited sensitivity to large view variations. MetaView achieves the best performance across all difficulties. Notably, on the hard subset, low-level metrics (PSNR, SSIM) become less indicative of perceptual differences, whereas our proposed DMD metric more clearly highlights the performance gap. Furthermore, MetaView consistently yields the best overall results on the RealEstate and Sekai datasets (Tab. 2), demonstrating strong robustness and generalization.

### 4.4 Ablation Study

We conduct ablations on the geometry tokens, the z-axis in RoPE and the pretrained backbone. To ensure efficiency, we conduct the following experiments on the DL3DV-medium test set. Fig. 5 presents qualitative results after ablating each component, while Tab. 3 reports the corresponding quantitative impact.

**Effect of Geometry Token.** We first ablate the implicit geometry priors by removing the geometry tokens. As shown in Fig. 5 (Col. 3), although the overall viewpoint orientation remains approximately correct, the model struggles to preserve complex relative relationships within the scene. For example, the signboard

**Table 3: Ablation study.** We conduct the training on DL3DV and testing on the DL3DV-medium. The effects of each component are reflected in the evaluation metrics.

| | PSNR↑ | SSIM↑ | LPIPS↓ | DMD↓ |
|---|---|---|---|---|
| Ours w/o geometry | 13.53 | 0.3538 | 0.2530 | 11.61 |
| Ours w/o $z$-axis | 12.49 | 0.3204 | 0.2900 | 14.45 |
| Ours w/ FLUX.1-Kontext | 13.90 | 0.3594 | 0.2571 | 10.22 |
| PE-Field (FLUX.1-Kontext) | 13.18 | 0.3435 | 0.2749 | 15.41 |
| Ours | **14.21** | **0.3765** | **0.2353** | **9.33** |

disappears and the fine-grained contours of objects have noticeable discrepancies. In the second row, the relative arrangement of objects deviates from the ground truth. Quantitatively, removing the geometry tokens results in performance degradation across all evaluation metrics. This decline aligns with our qualitative observations, as the absence of geometry priors weakens the model's ability to reason about fine-grained relative structure.

**Effect of $z$-axis.** We ablate the $z$-axis in RoPE by conditioning PE solely on pose and coordinates $(T, i, j)$. As shown in Fig. 5, removing the $z$-axis leads to a degradation in camera control accuracy. While the rotation remains correct, the viewpoint position exhibits a translation offset from the ground truth, reflecting the scale drifting issue. Incorporating the $z$-axis anchors scene scale and restores accurate control. Quantitatively (Tab. 3), removing it consistently degrades all metrics, particularly the low-level metrics PSNR and SSIM. This confirms that $z$-axis scale cues are essential for mitigating scale drifting issue and enabling scale-aware spatial reasoning.

**Effect of Backbone.** We adapt MetaView to the FLUX.1-Kontext backbone with the results summarized in Tab. 3 and Fig. 5. Using the same base model, MetaView outperforms PE-Field. While there is a moderate performance drop compared to our Qwen-Image-Edit variant, this degradation is primarily attributed to the inherent capability gap between the backbones. This confirms the effectiveness of our design and demonstrates the robustness of our framework in cross-model adaptation.

### 4.5 Analysis of DMD Metric

Low-level metrics fail in large-view monocular NVS, as extrapolated regions require plausible completion rather than strict pixel alignment. They are dominated by low-frequency alignment (Fig. 6), assigning higher scores to blurry artifacts (Output 2) and contradicting human perception. In contrast, our DMD aligns with human preference. In both cases, DMD correctly favors the visually superior Output 1. The numerical margin further demonstrates its discriminative ability in evaluating extrapolative synthesis.

Furthermore, DMD can numerically reflect the magnitude of viewpoint variations. As illustrated in Fig. 7, given the source image $X^{src}$, let $X^i$ denote a sequence of views with progressively increasing viewpoint changes. We compute the PSNR, SSIM, and LPIPS between each $X^i$ and $X^{src}$, along with our

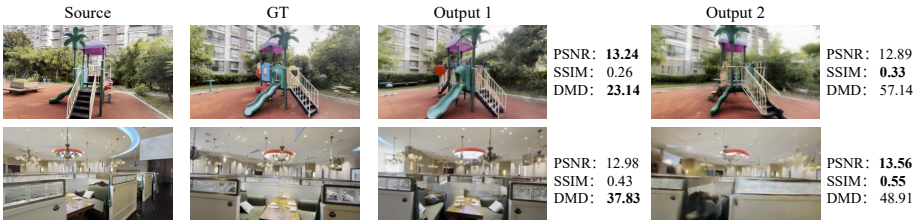


**Fig. 6: Comparison of metric reliability.** Low-frequency structural alignment dominates PSNR and SSIM, while our proposed DMD aligns with human preference.

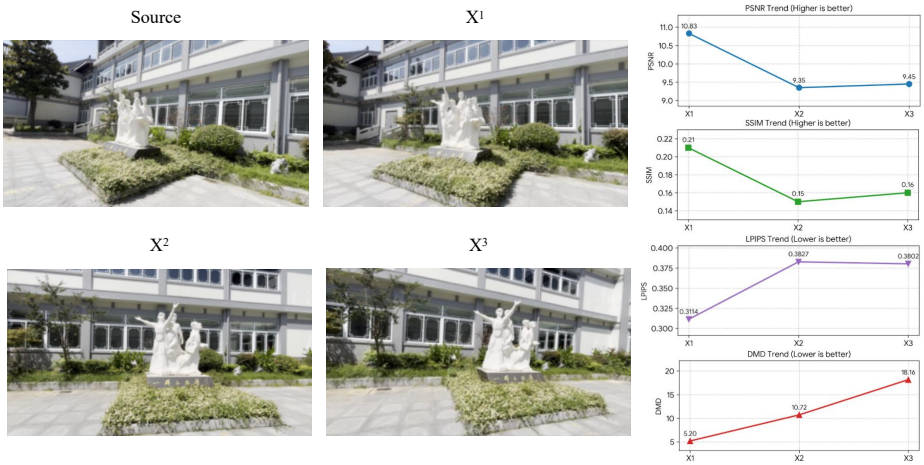


**Fig. 7: Trends of Different Metrics.** While traditional metrics (PSNR, SSIM, and LPIPS) fail to show a monotonic trend as the viewpoint changes, our proposed DMD exhibits monotonicity, accurately quantifying the discrepancy between views.

proposed metric $\mathrm{DMD}(X^i; X^i; X^{src})$. It can be observed that as the viewpoint variation increases, PSNR, SSIM, and LPIPS fail to exhibit a monotonic trend. In contrast, DMD demonstrates a monotonic variation, effectively capturing the viewpoint discrepancy between views at a numerical level.

In our implementation, when computing $\mathrm{DMD}(X^{src}; X^{gen}; X^{GT})$, the co-visibility threshold for calculating $P_{src} = \mathrm{UFM}(X^{src}; X^{GT})$ is set to 0.3, while the threshold for $P_{gen} = \mathrm{UFM}(X^{gen}; X^{GT})$ is set to 0.2.

### 4.6 Application

Our method also demonstrates strong generalization to diverse out-of-domain data. Since the pretrained backbone parameters remain frozen, the rich semantic priors acquired during large-scale pretraining are fully preserved. This design enables MetaView to maintain robust performance across a wide variety of


**Fig. 8: Application on open-domain data.** Our method preserves the rich semantic knowledge of the pretrained model, enabling generalization across diverse scenes.

scenes. As shown in Fig. 8, our approach produces consistent and accurate novel view synthesis results for scenes centered on humans and animals. Moreover, these examples span diverse visual styles, including cartoon, watercolor paintings, and AI-generated content. Such cross-domain robustness further highlights the practical applicability and downstream potential of MetaView.

## 5 Conclusion

In this paper, we present MetaView, a diffusion-based monocular novel view synthesis framework that achieves precise viewpoint control and large view generation without explicit 3D reconstruction. By bridging the gap between explicit geometric constraints and implicit generative modeling, MetaView effectively addresses the dual challenges of structural inconsistency and scale drifting. Specifically, we incorporate implicit geometry priors to guide fine-grained structural synthesis and preserve complex relative relationships, while introducing minimal yet essential 3D cues to anchor scene scale during generation. Extensive qualitative and quantitative experiments across multiple datasets demonstrate that our method consistently outperforms existing approaches. Moreover, MetaView exhibits strong generalization capability, including robustness to out-of-domain data. In future work, we will explore the spatial-aware generation in 4D scenes, enabling geometry-consistent in temporally evolving environments.

## 6 Acknowledgments

This research is supported by the MoE AcRF Tier 2 grant (MOE-T2EP20223-0001) and the MoE AcRF Tier 1 grant (RG14/22).

## References


1. Bai, Y., Li, H., Huang, Q.: Positional encoding field. arXiv preprint arXiv:2510.20385 (2025)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5855–5864 (2021)
3. Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., Leroy, V.: Must3r: Multi-view network for stereo 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1050–1060 (2025)
4. Cai, H., Cao, S., Du, R., Gao, P., Hoi, S., Hou, Z., Huang, S., Jiang, D., Jin, X., Li, L., et al.: Z-image: An efficient image generation foundation model with single-stream diffusion transformer. arXiv preprint arXiv:2511.22699 (2025)
5. Cao, C., Zhou, J., Li, S., Liang, J., Yu, C., Wang, F., Xue, X., Fu, Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025)
6. Chan, E.R., Nagano, K., Chan, M.A., Bergman, A.W., Park, J.J., Levy, A., Aittala, M., De Mello, S., Karras, T., Wetzstein, G.: Generative novel view synthesis with 3d-aware diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4217–4229 (2023)
7. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14124–14133 (2021)
8. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: European conference on computer vision. pp. 370–386. Springer (2024)
9. Debevec, P.E., Taylor, C.J., Malik, J.: Modeling and rendering architecture from photographs: a hybrid geometry-and image-based approach. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 11–20 (1996)
10. Deng, K., Ti, Z., Xu, J., Yang, J., Xie, J.: Vggt-long: Chunk it, loop it, align it–pushing vggt's limits on kilometer-scale long rgb sequences. arXiv preprint arXiv:2507.16443 (2025)
11. Duan, H., Yu, H.X., Chen, S., Fei-Fei, L., Wu, J.: Worldscore: A unified evaluation benchmark for world generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27713–27724 (2025)
12. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
13. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first International Conference on Machine Learning (2024)
14. He, X., Peng, C., Liu, Z., Wang, B., Zhang, Y., Cui, Q., Kang, F., Jiang, B., An, M., Ren, Y., et al.: Matrix-game 2.0: An open-source real-time and streaming interactive world model. arXiv preprint arXiv:2508.13009 (2025)

15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
16. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
17. Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.H., et al.: Vipe: Video pose engine for 3d geometric perception. arXiv preprint arXiv:2508.10934 (2025)
18. Huang, T., Zheng, W., Wang, T., Liu, Y., Wang, Z., Wu, J., Jiang, J., Li, H., Lau, R., Zuo, W., et al.: Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. ACM Transactions on Graphics (TOG) **44**(6), 1–15 (2025)
19. Jin, H., Jiang, H., Tan, H., Zhang, K., Bi, S., Zhang, T., Luan, F., Snavely, N., Xu, Z.: Lvsm: A large view synthesis model with minimal 3d inductive bias. In: The Thirteenth International Conference on Learning Representations
20. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
21. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
22. LEVOY, M.: Light field rendering. In: Computer Graphics Proceedings, Annual Conference Series, Proc. SIGGRAPH'96 (New Orleans), August 1996. ACM SIGGRAPH (1996)
23. Li, R., Yi, B., Liu, J., Gao, H., Ma, Y., Kanazawa, A.: Cameras as relative positional encoding. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
24. Li, Z., Li, C., Mao, X., Lin, S., Li, M., Zhao, S., Li, X., Feng, Y., Sun, J., Li, Z., et al.: Sekai: A video dataset towards world exploration. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track
25. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
26. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22160–22169 (2024)
27. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations
28. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: 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)
29. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations
30. Maggio, D., Lim, H., Carlone, L.: Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. arXiv preprint arXiv:2505.12549 (2025)
31. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision. pp. 405–421. Springer (2020)

32. Miyato, T., Jaeger, B., Welling, M., Geiger, A.: Gta: A geometry-aware attention mechanism for multi-view transformers. In: The Twelfth International Conference on Learning Representations
33. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG) **41**(4), 1–15 (2022)
34. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning. pp. 16784–16804 (2022)
35. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
36. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
37. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
38. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6121–6132 (2025)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
41. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
42. Sargent, K., Li, Z., Shah, T., Herrmann, C., Yu, H.X., Zhang, Y., Chan, E.R., Lagun, D., Fei-Fei, L., Sun, D., 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)
43. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
44. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: European conference on computer vision. pp. 501–518. Springer (2016)
45. Seo, J., Fukuda, K., Shibuya, T., Narihira, T., Murata, N., Hu, S., Lai, C.H., Kim, S., Mitsufuji, Y.: Genwarp: Single image to novel views with semantic-preserving generative warping. Advances in Neural Information Processing Systems **37**, 80220–80243 (2024)
46. Sinha, S., Steedly, D., Szeliski, R.: Piecewise planar stereo for image-based rendering. In: 2009 International Conference on Computer Vision. pp. 1881–1888 (2009)
47. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)

48. Sun, C., Sun, M., Chen, H.T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5459–5469 (2022)
49. Sun, W., Zhang, H., Wang, H., Wu, J., Wang, Z., Wang, Z., Wang, Y., Zhang, J., Wang, T., Guo, C.: Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. arXiv preprint arXiv:2512.14614 (2025)
50. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)
51. Tang, Z., Fan, Y., Wang, D., Xu, H., Ranjan, R., Schwing, A., Yan, Z.: Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5283–5293 (2025)
52. Team, R., Gao, Z., Wang, Q., Zeng, Y., Zhu, J., Cheng, K.L., Li, Y., Wang, H., Xu, Y., Ma, S., et al.: Advancing open-source world models. arXiv preprint arXiv:2601.20540 (2026)
53. Van Hoorick, B., Chen, D., Iwase, S., Tokmakov, P., Irshad, M.Z., Vasiljevic, I., Gupta, S., Cheng, F., Zakharov, S., Guizilini, V.C.: Anyview: Synthesizing any novel view in dynamic scenes. arXiv preprint arXiv:2601.16982 (2026)
54. Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., Dave, A., Zheng, C., Vondrick, C.: Generative camera dolly: Extreme monocular dynamic novel view synthesis. In: European Conference on Computer Vision. pp. 313–331. Springer (2024)
55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
56. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
57. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20697–20709 (2024)
58. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: $\pi$3: Scalable permutation-equivariant visual geometry learning. arXiv e-prints pp. arXiv–2507 (2025)
59. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
60. Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V.A.: Nerf–: Neural radiance fields without known camera parameters (2021)
61. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
62. Xu, G., Wang, X., Ding, X., Yang, X.: Iterative geometry encoding volume for stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21919–21928 (2023)
63. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21924–21935 (2025)
64. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European conference on computer vision (ECCV). pp. 767–783 (2018)

65. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4578–4587 (2021)
66. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
67. Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A.: Mip-splatting: Alias-free 3d gaussian splatting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19447–19456 (2024)
68. Zhang, J., Herrmann, C., Hur, J., Jampani, V., Darrell, T., Cole, F., Sun, D., Yang, M.H.: Monst3r: A simple approach for estimating geometry in the presence of motion. arXiv preprint arXiv:2410.03825 (2024)
69. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: 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)
70. Zhang, Y., Keetha, N.V., Lyu, C., Jhamb, B., Chen, Y., Qiu, Y., Karhade, J., Jha, S., Hu, Y., Ramanan, D., et al.: Ufm: A simple path towards unified dense correspondence with flow. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
71. Zhou, J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12405–12414 (2025)
72. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: learning view synthesis using multiplane images. ACM Transactions on Graphics (TOG) **37**(4), 1–12 (2018)