Repurposing Geometric Foundation Models for Multi-view Diffusion

Wooseok Jang[1], Seonghu Jeon[1], Jisang Han[1], Jinhyeok Choi[1], Minkyung Kwon[1], Seungryong Kim[1], Saining Xie[2], and Sainan Liu[3]

[1]KAIST AI [2]New York University [3]Intel Labs

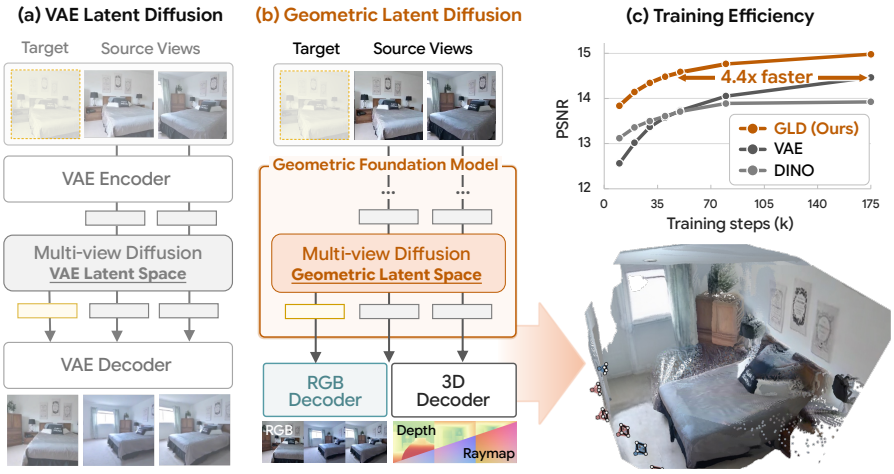


Fig. 1: **(a)** Prior multi-view diffusion operates in VAE latent space. **(b) Geometric Latent Diffusion (GLD)** diffuses in the feature space of geometric foundation models [25,45], enabling both RGB and geometry decoding. **(c)** GLD converges **4.4**× faster than VAE [35] and DINO [52].

Abstract. The latent space of diffusion models fundamentally determines their learning efficiency and generation quality. While recent advances in the latent space have driven substantial progress in single-image generation, the optimal latent space for novel view synthesis (NVS) remains largely unexplored. In particular, NVS requires geometrically consistent generation across viewpoints, but existing approaches typically operate in a view-independent latent space. In this paper, we propose **Geometric Latent Diffusion (GLD)**, a framework that repurposes the feature space of a geometric foundation model as the latent space for multi-view diffusion. We show that the features of the geometric foundation model not only support high-fidelity RGB reconstruction but also encode strong cross-view geometric correspondences, providing a well-suited latent space for NVS. Through experiments, GLD outperforms both VAE and RAE on 2D image quality and 3D consistency metrics, accelerating training by more than **4.4**× compared to the VAE latent space. Notably, GLD remains competitive with state-of-the-art methods that leverage large-scale text-to-image pretraining, despite training its diffusion model from scratch without such generative pretraining.

Keywords: Novel View Synthesis · Latent Diffusion · Geometric Foundation Models

1 Introduction

Diffusion models [13,42] have become the dominant framework for image synthesis. Transitioning from pixel space to the variational auto-encoders (VAEs) latent space [22,32,35], and further to semantically structured representations [39,43,52], has shown that the latent space significantly influences generation quality and training efficiency. However, these insights have been drawn exclusively from 2D image generation, and the design of effective latent spaces for **geometry-aware generation tasks** remains largely unexplored.

Novel view synthesis (NVS), which predicts unseen viewpoints consistent with an underlying 3D scene [20,38], is a representative geometry-aware generation task. Unlike single-image generation, this task requires maintaining coherent spatial structure across views and geometrically plausible completion of occluded regions. While early generative approaches [11,40] have demonstrated photorealistic image quality, they often prioritize appearance over geometric consistency, leading to geometrically inconsistent outputs. To address this, recent diffusion-based NVS methods often leverage external geometry conditioning, such as depth-based warping [4,20,38,51]. However, these approaches still rely on latent spaces originally designed for single-image synthesis models, such as the 2D VAEs [35]. This raises a fundamental question: **can we leverage a latent space in which geometric structure is already encoded, rather than injecting or supervising it externally?**

In this work, we propose **Geometric Latent Diffusion (GLD)** that utilizes the feature space of geometric foundation models [25,45,49] such as VGGT [45] or Depth Anything 3 (DA3) [25] as a **latent space for multi-view diffusion models**. We first show that the features of a geometric foundation model support high-fidelity, view-consistent RGB reconstruction, enabling the diffusion process to operate directly on geometry-aware representations for NVS. These models typically produce a hierarchy of multi-level features to reconstruct 3D geometries. To ensure computational efficiency, rather than diffusing the entire multi-level features, we identify an optimal boundary layer level for explicit synthesis. Deeper-layer features beyond this boundary are naturally derived by propagating through the frozen backbone, while shallower features are generated via a cascaded scheme to ensure cross-level alignment. By training in this geometry-informed latent space, GLD leverages the rich geometric structure, which provides the necessary grounding for generating view-consistent images. Furthermore, since our framework operates natively on the geometric foundation model's features, synthesized latents can be directly decoded into geometric predictions (*e.g.*, depth maps and camera poses) without additional training.

Through extensive experiments on both in-domain and zero-shot benchmarks, GLD achieves superior pixel-level fidelity and cross-view 3D consistency compared with VAE [35] and RAE [52] baselines, accelerating training conver-

gence by over 4.4×. Although our diffusion model is trained from scratch on small datasets, GLD remains competitive with state-of-the-art methods [4, 11, 21, 23, 29], fine-tuned from large-scale text-to-image models. Moreover, zero-shot depth and 3D point clouds decoded from synthesized latents exhibit strong global consistency. These results validate that our GLD framework effectively integrates generative modeling with a geometry-informed latent representation.

2 Related Work

Novel View Synthesis with Diffusion Models. Classical geometry-based approaches to novel view synthesis [16, 30] produce photorealistic renderings but require dense multi-view captures and costly per-scene optimization. Recent multi-view diffusion models [4, 11, 18, 20, 29, 38, 51, 53] alleviate these constraints by leveraging generative priors to synthesize novel views from sparse inputs. However, these methods operate in pixel or VAE latent spaces that lack cross-view geometric structure, placing a substantial burden on the model to implicitly discover geometric correspondences [21]. We instead train multi-view diffusion models in a latent space that already encodes this structure.

Latent Spaces for Diffusion Models. Latent diffusion models (LDMs) [35] have advanced image synthesis by operating in a compressed VAE [17] latent space, but it lacks rich structural priors. RAE [43, 52] and SVG [39] show that frozen semantic encoders [31, 41, 44] can be paired with lightweight decoders for high-fidelity reconstruction, and that diffusing in this semantic space yields faster convergence and improved generation quality. However, these advances target single-image generation, leaving open the question of how to design latent spaces for geometry-aware generation tasks. Recent works address this by training dedicated autoencoders that jointly encode appearance and geometry, for single-image [19] and text-to-3D [50] generation. Our work instead repurposes the feature space of an existing geometric foundation model [25] as the latent space for diffusion, providing the model with cross-view geometric priors.

Geometric Foundation Models. Geometric foundation models have introduced a paradigm shift in 3D vision, moving from optimization-based feature matching [36] to purely feed-forward scene understanding. Building on the pairwise formulation of DUSt3R [47], recent models [15, 25, 45, 49] have enabled feed-forward dense 3D reconstruction from arbitrary unposed views, jointly predicting camera parameters and depth maps. While recent analyses reveal that the internal representations of these networks encode strong geometric correspondences [12], their utility has been largely limited to discriminative tasks. We bridge this gap by showing that the feature space of a geometric foundation model [25] can serve as an effective latent space for novel view synthesis.

3 Preliminaries

Representation Autoencoder. Representation Autoencoder (RAE) [43, 52] replaces the conventional VAE [17] in latent diffusion [35] with a pretrained,

frozen vision encoder [31,44] $\mathcal{E}(\cdot)$ and a trainable decoder $\mathcal{D}(\cdot)$, directly adopting the encoder's feature space as the diffusion latent space.

Specifically, the decoder is trained to reconstruct RGB images from features in this representation space, showing that these features are not only semantically rich but also sufficient for high-fidelity reconstruction. Formally, given a single-view image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the height and width, respectively, the encoder extracts a tokenized feature representation

$$\mathbf{F} = \mathcal{E}(\mathcal{I}) \in \mathbb{R}^{T\times C}, \tag{1}$$

where T is the token sequence length and C is the channel dimension. RAE further shows that diffusion models can be trained directly in this representation space, yielding faster convergence and stronger generative performance than training in conventional VAE latent spaces [17]. During generation, the diffusion model synthesizes $\tilde{\mathbf{F}}$, and the synthesized image $\tilde{\mathcal{I}}$ is then obtained by decoding the synthesized feature: $\tilde{\mathcal{I}} = \mathcal{D}(\tilde{\mathbf{F}})$.

Geometric Foundation Models. Recent foundation models for geometry [25, 45, 49] typically consist of a Vision Transformer (ViT) [8] encoder $\mathcal{E}_{\text{geo}}(\cdot)$ and a DPT-based geometric decoder $\mathcal{D}_{\text{geo}}(\cdot)$. To process multi-view inputs, these architectures often incorporate 3D attention in addition to standard intra-image self-attention, which enables joint reasoning across multiple frames. Given multi-view images $\mathbf{I} \in \mathbb{R}^{V\times H\times W\times 3}$, where V is the number of input views, the encoder extracts multi-view feature sequences at L levels:

$$\{\mathbf{F}_l\}_{l=0}^{L-1} = \mathcal{E}_{\text{geo}}(\mathbf{I}), \tag{2}$$

where $\mathbf{F}_l \in \mathbb{R}^{V\times T\times C}$ denotes the multi-view feature sequence at level l, with T the token sequence length and C the channel dimension. The geometric decoder then aggregates these multi-level features to produce dense geometric predictions, such as depth or camera parameters:

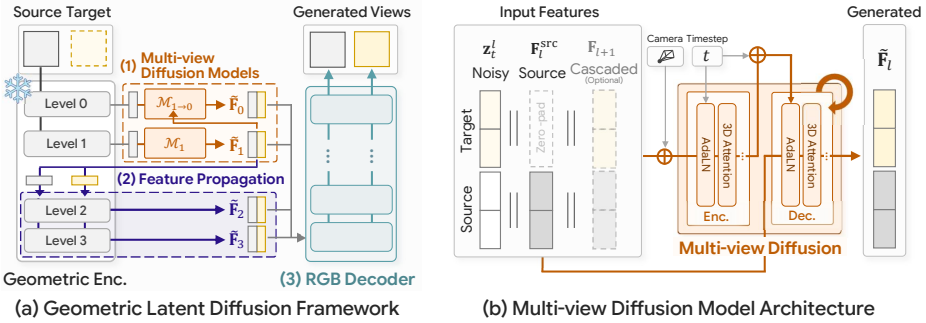
$$\mathbf{G} = \mathcal{D}_{\text{geo}}(\{\mathbf{F}_l\}_{l=0}^{L-1}), \tag{3}$$

where $\mathbf{G}$ denotes a set of geometric predictions for each input view.

4 Method

4.1 Overview

Our goal is to harness the feature space of geometric foundation models [25,45,49] as the latent space for multi-view diffusion, enabling high-fidelity novel view synthesis (NVS). Specifically, we adopt the Depth Anything 3 (DA3) [25] as our primary backbone, which extracts features across $L = 4$ intermediate levels. We also explore VGGT [45] as an additional backbone in Appendix C.1. Given a set of N source images $\mathbf{I}^{\text{src}} \in \mathbb{R}^{N\times H\times W\times 3}$ with camera poses $\mathbf{P}^{\text{src}}$, and M target camera poses $\mathbf{P}^{\text{tgt}}$, we seek to synthesize the corresponding target views $\tilde{\mathbf{I}}^{\text{tgt}} \in \mathbb{R}^{M\times H\times W\times 3}$. Rather than operating directly in pixel space or VAE space [35], our



(a) Geometric Latent Diffusion Framework (b) Multi-view Diffusion Model Architecture

Fig. 2: Overview of the proposed framework. (a) Our NVS framework consists of three stages: **(1) multi-view diffusion models** synthesize features up to boundary layer k=1 via cascaded generation, **(2) feature propagation** derives deeper features through the frozen DA3 encoder, and **(3) the RGB decoder** maps the full multi-level features to target views. **(b)** The detailed architecture of a multi-view diffusion model.

framework generates multi-view, multi-level features from geometric foundation models [25, 45, 49], which are subsequently decoded into the target views.

Because the geometric foundation model's 3D attention jointly encodes source and target views, the resulting features are inherently coupled. We therefore generate both source and target features across all levels, denoted as $\{\tilde{\mathbf{F}}_l^{\text{src}}\}_{l=0}^{L-1}$ with $\tilde{\mathbf{F}}_l^{\text{src}} \in \mathbb{R}^{N\times T\times C}$, and $\{\tilde{\mathbf{F}}_l^{\text{tgt}}\}_{l=0}^{L-1}$ with $\tilde{\mathbf{F}}_l^{\text{tgt}} \in \mathbb{R}^{M\times T\times C}$. At each level l, the joint feature $\tilde{\mathbf{F}}_l \in \mathbb{R}^{(N+M)\times T\times C}$ is formed by concatenating the source and target features along the view dimension, yielding $\tilde{\mathbf{F}}_l = [\tilde{\mathbf{F}}_l^{\text{src}}, \tilde{\mathbf{F}}_l^{\text{tgt}}]$ with concatenation operator $[\cdot,\cdot]$. Finally, a dedicated RGB decoder $\mathcal{D}_{\text{rgb}}(\cdot)$ maps the complete synthesized feature set back to the pixel space to render the target views via $\tilde{\mathbf{I}}^{\text{tgt}} = \mathcal{D}_{\text{rgb}}(\{\tilde{\mathbf{F}}_l^{\text{tgt}}\}_{l=0}^{L-1})$. The target geometry $\tilde{\mathbf{G}}^{\text{tgt}}$ is also decoded such that $\tilde{\mathbf{G}}^{\text{tgt}} = \mathcal{D}_{\text{geo}}(\{\tilde{\mathbf{F}}_l^{\text{tgt}}\}_{l=0}^{L-1})$.

To this end, as illustrated in Fig. 2, **Geometric Latent Diffusion (GLD)** framework employs a three-stage pipeline. First, § 4.2 validates the reconstruction capacity of the geometric feature space by training $\mathcal{D}_{\text{rgb}}(\cdot)$ to decode multi-level features into RGB images. Second, to avoid the substantial cost of diffusing all L feature levels, § 4.3 identifies the optimal boundary layer k. Since deeper features can be obtained by propagating the boundary feature $\tilde{\mathbf{F}}_k$ through $\mathcal{E}_{\text{geo}}(\cdot)$, we only require explicit synthesis up to this boundary layer. Finally, § 4.4 employs a cascaded scheme to synthesize shallower features from $\tilde{\mathbf{F}}_k$ to ensure cross-level alignment.

4.2 Validating the Reconstruction Capability of Geometric Features

To validate the suitability of DA3's feature space for generative modeling, we first verify that its features can be decoded into high-fidelity images. We train a ViT-based decoder $\mathcal{D}_{\text{rgb}}(\cdot)$ to reconstruct RGB images from the multi-level features $\{\mathbf{F}_l\}_{l=0}^{L-1}$ extracted by the frozen encoder $\mathcal{E}_{\text{geo}}(\cdot)$. To ensure $\mathcal{D}_{\text{rgb}}(\cdot)$ effectively leverages the full signal, we introduce a level-wise dropout strategy during training. By randomly masking individual levels in $\{\mathbf{F}_l\}_{l=0}^{L-1}$, we force the

Fig. 3: Reconstruction Results. We visualize the reconstructed images from $\mathcal{D}_{\text{rgb}}$.

Table 1: Reconstruction Fidelity. We evaluate our RGB decoder on 4,000 samples from the Re10K [54] test set.

Metrics	Decoder ($\mathcal{E}_{\text{rgb}}$)
PSNR ↑	35.41
SSIM ↑	0.960
LPIPS ↓	0.019

decoder to reconstruct from partial inputs, improving its robustness. As shown in Tab. 1 and Fig. 3, $\mathcal{D}_{\text{rgb}}(\cdot)$ successfully recovers the input images with high fidelity while preserving fine-grained details. These results show that the DA3 feature space is suitable as the latent space for our diffusion process. Further details and comparison with other baselines are available in § 5.4.

4.3 Determining the Boundary Layer for Explicit Synthesis

While the DA3 feature space provides a sufficiently expressive latent for high-fidelity image reconstruction, explicitly synthesizing the full multi-level set $\{\mathbf{F}_l\}_{l=0}^{3}$ is computationally prohibitive. Since deeper features ($l > k$) can be deterministically derived by propagating a shallower feature $\tilde{\mathbf{F}}_k$ through the frozen layers of $\mathcal{E}_{\text{geo}}$, we only require explicit synthesis up to an optimal boundary k. To identify this boundary, we first train four independent diffusion models $\{\mathcal{M}_l\}_{l=0}^{3}$, each dedicated to synthesizing the target feature at a specific level l. We then perform a comparative evaluation by varying the synthesis boundary $k \in \{0, 1, 2, 3\}$ to identify the shallowest boundary sufficient for high-quality NVS.

Multi-view Diffusion Architecture and Training. We adopt the DiT$^{\text{DH}}$ [46] architecture from RAE [52] and train it with a flow-matching objective [27] to synthesize the joint feature map $\tilde{\mathbf{F}}_l$ for all $V = N + M$ views. As illustrated in Fig. 2, we incorporate 3D self-attention [11] with PRoPE [24] and condition on Plücker ray embeddings to enforce geometric consistency across views.

Each model $\mathcal{M}_l$ is conditioned on the source-only features $\mathbf{F}_l^{\text{src}}$, extracted by the frozen DA3 encoder from the N source images alone, by concatenating them with the noisy latent $\mathbf{z}_t$ along the channel dimension.

Note that $\mathbf{F}_l^{\text{src}}$ is extracted without access to target views, whereas the source portion of the full joint feature $\mathbf{F}_l$ is influenced by 3D attention over all V views. Because the decoder and downstream stages require features from all views, we design the model to jointly generate $\mathbf{F}_l$ rather than generate only the target views.

Boundary Layer Evaluation To identify the optimal boundary, we assess how each level k contributes to the generation by providing the decoder $\mathcal{D}_{\text{rgb}}$ with a complete multi-level set $\{\tilde{\mathbf{F}}_l\}_{l=0}^{3}$. For a given boundary k, we explicitly synthesize the features up to that level ($l \leq k$), $\{\tilde{\mathbf{F}}_l\}_{l=0}^{k}$, using the corresponding set of independently trained models $\{\mathcal{M}_l\}_{l=0}^{k}$. The remaining deeper levels ($l > k$) are then deterministically derived by passing $\tilde{\mathbf{F}}_k$ through the frozen layers of $\mathcal{E}_{\text{geo}}$.

Table 2: Boundary Selection. We evaluate NVS performance by varying the synthesis boundary k. Explicitly synthesizing up to level $l = 1$ (Boundary $k = 1$) using $\mathcal{M}_0$ and $\mathcal{M}_1$ provides the best performance across all metrics.

Synthesis Configuration		RGB Metrics			Depth Metrics		
Boundary k	Models $\{\mathcal{M}_l\}_{l=0}^{k}$	PSNR ↑	SSIM ↑	LPIPS ↓	AbsRel ↓	RMSE ↓	$\delta < 1.25$ ↑
$k = 0$	$\{\mathcal{M}_0\}$	12.55	0.323	0.579	0.267	0.400	0.641
$k = 1$	$\{\mathcal{M}_0, \mathcal{M}_1\}$	**13.61**	**0.366**	**0.555**	**0.191**	**0.311**	**0.744**
$k = 2$	$\{\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2\}$	13.35	0.355	0.566	0.254	0.393	0.659
$k = 3$	$\{\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3\}$	13.35	0.355	0.567	0.260	0.402	0.647

As shown in Tab. 2, synthesizing up to level 1 achieves superior NVS performance. Shifting the boundary from level 0 to level 1 improves both RGB quality and geometric accuracy, suggesting that level 1 provides a more effective latent representation for synthesis. Conversely, using deeper levels (2 or 3) as the boundary leads to a consistent degradation in metrics, likely due to the loss of fine-grained spatial details in abstract feature spaces. Consequently, we fix level 1 as the synthesis boundary k for our full framework, as illustrated in Fig. 2(a). Further analysis of this selection is provided in § 5.5.

4.4 Cascaded Feature Generation

Based on the evaluation in § 4.3, we fix the synthesis boundary at level 1 and generate the corresponding multi-level set. While level 0 can be synthesized independently, generating them separately causes misalignment across the feature hierarchy. To provide the coherent input required by the decoder, we instead employ a cascaded model $\mathcal{M}_{1\to0}$ that synthesizes level 0 conditioned on the generated level 1 latent, which is illustrated Fig. 2(b). $\mathcal{M}_{1\to0}$ shares the same architecture and training configuration as $\mathcal{M}_0$.

To handle the imperfect latents encountered during inference, we train $\mathcal{M}_{1\to0}$ by conditioning on a noisy version of the ground-truth $\mathbf{F}_1$. This strategy improves the model's robustness and ensures $\tilde{\mathbf{F}}_0$ is anchored to $\tilde{\mathbf{F}}_1$, providing the alignment the decoder requires. A quantitative validation of this cascaded approach over independent generation is provided in § 5.6.

5 Experiments

5.1 Setup

Dataset. GLD is trained from scratch on four datasets, RealEstate10K [54] (Re10K), DL3DV [26], HyperSim [34], and TartanAir [48]. Each training sample consists of $V = 8$ views, where 1 to 4 views are randomly selected as source views while the rest are masked as targets. For the main evaluation, we evaluate on two in-domain benchmarks, Re10K and DL3DV, which overlap with the training distribution, and on one out-of-domain object-centric benchmark, Mip-NeRF 360 [3], to evaluate generalization to unseen scene types. We use $N = 2$ source views and measure performance on the target views across 200 samples per dataset. Additional details are provided in Appendix A.3.

Table 3: Quantitative comparison of different methods across in-domain [26, 54] and out-of-domain [3] datasets on 2D and 3D metrics. **Bold** and underlined values indicate the best and second-best results, respectively. We use reproduced CAT3D† [21].

Methods	From Scratch	2D Metrics			3D Metrics				
		PSNR ↑	SSIM ↑	LPIPS ↓	ATE ↓	RPE_r ↓	RPE_t ↓	Reproj ↓	MEt3R ↓
RealEstate10K [54]									
MVGenMaster [4]	✗	15.226	0.588	0.456	0.282	**6.42**	0.526	0.664	0.339
Matrix3D [29]	✗	14.490	0.580	0.448	0.413	8.93	0.638	0.666	0.344
CAMEO [21]	✗	13.800	0.561	0.522	0.446	12.93	0.790	**0.661**	0.344
NVComposer [23]	✗	11.140	0.418	0.649	0.829	42.85	1.435	0.860	0.457
CAT3D† [11]	✗	13.350	0.527	0.561	0.496	17.49	0.941	0.719	0.361
DINO [31]	✓	15.638	0.601	0.448	0.345	15.59	0.719	0.721	**0.319**
VAE [35]	✓	15.656	0.606	0.456	0.278	8.68	0.552	0.681	0.375
GLD (Ours)	✓	**16.362**	**0.630**	**0.431**	**0.211**	7.07	**0.444**	0.673	0.328
DL3DV [26]									
MVGenMaster [4]	✗	14.565	0.442	0.460	0.281	6.93	0.592	0.637	**0.375**
Matrix3D [29]	✗	13.330	0.396	0.451	0.459	9.65	0.850	0.667	0.394
CAMEO [21]	✗	12.320	0.371	0.567	1.143	24.76	2.149	0.706	0.404
NVComposer [23]	✗	10.518	0.273	0.646	1.810	55.59	3.098	0.852	0.517
CAT3D† [11]	✗	11.820	0.335	0.594	1.346	31.73	2.473	0.746	0.435
DINO [31]	✓	14.345	0.411	0.471	0.546	13.12	1.050	0.708	0.410
VAE [35]	✓	14.725	0.446	0.476	0.589	15.00	1.116	0.674	0.407
GLD (Ours)	✓	**15.499**	**0.468**	**0.438**	**0.209**	**5.75**	**0.466**	**0.612**	0.378
Mip-NeRF 360 [3] (Out-of-domain)									
MVGenMaster [4]	✗	14.170	**0.304**	0.511	**0.320**	**10.92**	**0.587**	0.676	0.402
Matrix3D [29]	✗	13.970	0.284	**0.483**	0.548	13.63	0.948	0.646	0.422
CAMEO [21]	✗	11.900	0.250	0.629	1.623	48.75	3.008	0.684	**0.395**
NVComposer [23]	✗	12.525	0.217	0.637	1.622	54.26	2.703	0.767	0.526
CAT3D† [11]	✗	11.310	0.214	0.653	1.722	54.65	3.171	0.724	0.453
DINO [31]	✓	13.718	0.267	0.542	0.949	27.57	1.720	0.707	0.444
VAE [35]	✓	13.942	0.274	0.548	1.221	35.34	2.200	0.674	0.449
GLD (Ours)	✓	**14.542**	0.288	0.504	0.589	15.97	1.071	**0.630**	0.406

Training Detail. Our diffusion model operates in the latent space of DA3-Base [25]. We train the diffusion model using AdamW [28] with a fixed learning rate of 5×10^{-5}, a batch size of 48, and EMA decay of 0.9995. We apply 10% dropout to camera embeddings for classifier-free guidance [14], with a CFG scale of 1.5 at inference. For each sample, the training resolution is randomly chosen from (504×504), (504×378), (504×336), and (504×280), matching the set of resolutions used to train DA3. GLD is trained on 8 B200 GPUs for 175k iterations. Additional details are provided in Appendix A.1.

Baselines. We compare GLD against two categories of baselines. The first category comprises general-purpose visual encoders, such as VAE [35] and DINO [31], to assess whether the latent space of DA3 [25] is better suited for novel view synthesis than general-purpose visual representations. We additionally evaluate VGGT [45] as an alternative geometric foundation model backbone to further examine the effectiveness of geometry-aware latent spaces for NVS; these results are provided in Appendix C.1. For VAE, we use the Stable Diffusion encoder-decoder [35]. For DINO, we adopt DINOv2 ViT-B/14 with registers [6, 31] as the encoder and train a decoder from scratch. In all cases, the diffusion model is trained from scratch for the same number of iterations using the same architecture as GLD. Additional implementation details are provided in the Appendix A.2.

The second category comprises state-of-the-art diffusion-based NVS methods, including MVGenMaster [4], Matrix3D [29], CAMEO [21], NVComposer [23], and CAT3D† [11][1]. These models typically leverage powerful generative priors by fine-tuning from large-scale pre-trained weights. Note that CAMEO and CAT3D† are trained exclusively on the Re10K dataset, whereas the remaining methods incorporate both scene-centric and object-centric data during training.

Evaluation Protocol. We evaluate the 2D image fidelity of generated target views using standard NVS metrics: PSNR, SSIM, and LPIPS. To assess the 3D geometric consistency of generated views, we further incorporate camera estimation errors, reprojection error [9], and MEt3R [2]. Specifically, for camera errors, we extract camera poses from the generated views using an external estimator [45] to compute the Absolute Trajectory Error (ATE), and Relative Pose Errors for rotation (RPE_r) and translation (RPE_t). These camera errors explicitly evaluate condition fidelity by measuring how accurately the generated images adhere to the target pose conditioning. Furthermore, reprojection error and MEt3R quantify the underlying 3D geometric consistency across the generated images. Reprojection error measures the spatial re-alignment accuracy of reconstructed 3D points, while MEt3R evaluates multi-view consistency using projected feature similarity.

5.2 Quantitative Results

2D Metrics. We evaluate image synthesis quality on two in-domain datasets (Re10K and DL3DV) and one zero-shot, out-of-domain benchmark (Mip-NeRF 360). First, we compare our performance against VAE and DINO encoder baselines. As shown in Tab. 3, our method consistently outperforms both baselines in PSNR, SSIM, and LPIPS across all benchmarks. The consistent gains confirm that the DA3 feature space provides a more suitable latent representation for novel view synthesis (NVS) than general-purpose visual encoders.

Second, we evaluate GLD against state-of-the-art NVS methods that leverage massive diffusion priors via large-scale text-to-image (T2I) pretraining. Despite being trained from scratch on smaller datasets, GLD surpasses all baselines across all 2D metrics on both in-domain benchmarks. For the out-of-domain evaluation, which consists mainly of object-centric samples, GLD still achieves state-of-the-art PSNR and highly competitive results across the remaining 2D metrics. This generalization is particularly notable given that GLD is trained exclusively on scene-level data, whereas competing baselines [4, 21, 23] incorporate object-centric datasets [7, 33] during fine-tuning.

3D Metrics. We next evaluate cross-view geometric consistency. As shown in Tab. 3, GLD consistently outperforms the VAE and DINO baselines across most 3D metrics on every benchmark. The most substantial gains are in pose accuracy, where GLD achieves up to a 2.8× lower ATE and a 2.6× lower RPE

[1] Since the official implementation of CAT3D is unavailable, we use the model and checkpoint reproduced in CAMEO [21].

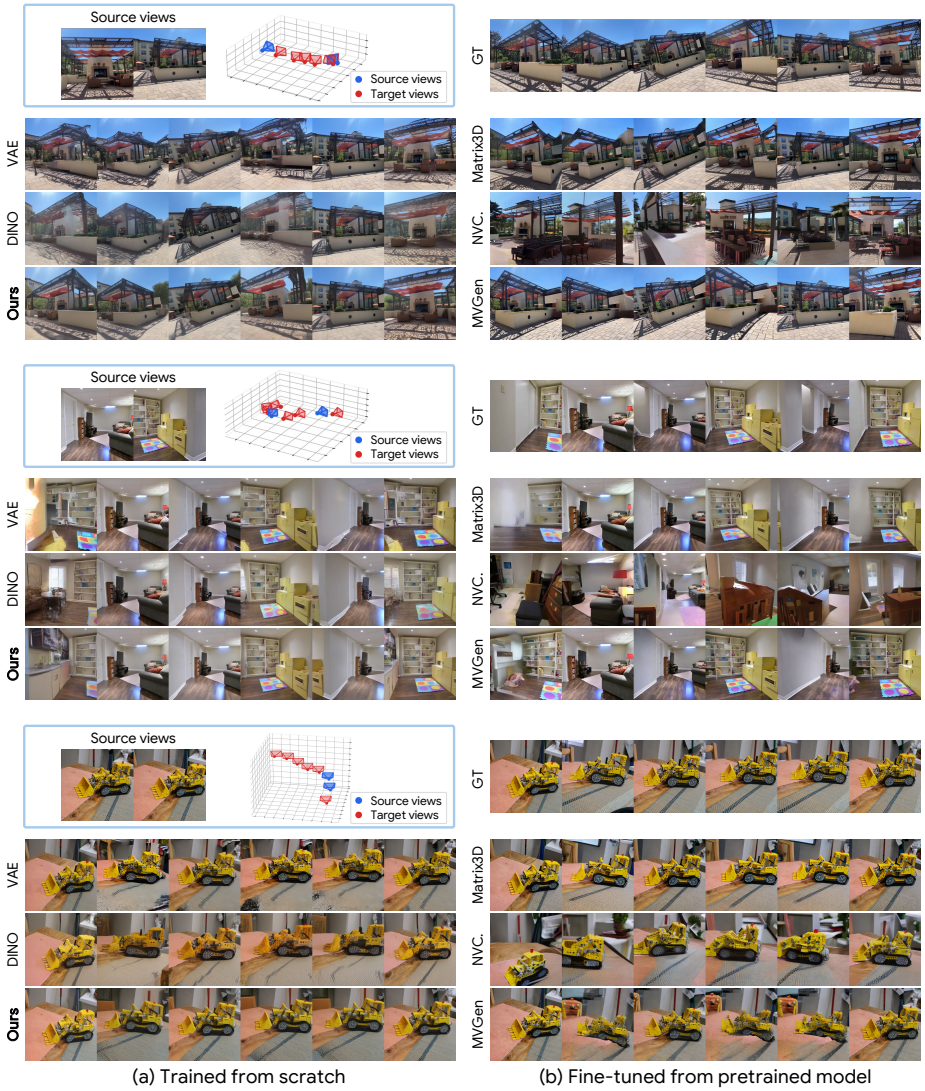


Fig. 4: Qualitative results. We compare the rendering quality of target views given two source views. The rows from top to bottom correspond to the Re10K, DL3DV, and out-of-domain Mip-NeRF 360 datasets, respectively.

compared to the baselines. Reprojection error and MEt3R also show consistent improvements across most evaluation settings. These significant improvements compared to VAE and DINO latent spaces demonstrate that GLD yields superior 3D consistency and notably accurate adherence to the target pose.

When compared to most fine-tuned methods, GLD exhibits significantly superior 3D consistency across in-domain benchmarks. This performance gap highlights a fundamental limitation of relying solely on generic T2I priors for NVS. Compared to MVGenMaster, GLD achieves superior performance on DL3DV,

Table 4: Ablation study across different numbers of source images. We report results with N=1 and N=4 input source views on Re10K and DL3DV. **Bold** values indicate best results.

Source Views	Dataset	Methods	2D Metrics			3D Metrics				
			PSNR ↑	SSIM ↑	LPIPS ↓	ATE ↓	RPE_r ↓	RPE_t ↓	Reproj ↓	MEt3R ↓
N=1	Re10K	DINO [31]	13.49	0.543	**0.541**	0.442	22.95	0.972	0.735	**0.301**
		VAE [35]	12.99	0.519	0.566	0.371	12.47	0.718	0.689	0.387
		GLD (Ours)	**13.50**	**0.552**	**0.541**	**0.267**	**8.42**	**0.539**	**0.673**	0.306
	DL3DV	DINO [31]	12.80	0.368	**0.539**	0.743	20.14	1.405	0.738	0.418
		VAE [35]	12.47	0.360	0.569	0.880	25.93	1.828	0.702	0.420
		GLD (Ours)	**13.03**	**0.385**	0.543	**0.237**	**6.14**	**0.512**	**0.619**	**0.375**
N=4	Re10K	DINO [31]	17.73	0.651	0.365	0.221	7.92	0.482	0.685	**0.327**
		VAE [35]	18.68	0.695	0.348	0.200	**6.26**	0.407	0.655	0.347
		GLD (Ours)	**19.00**	**0.696**	**0.327**	**0.182**	6.69	**0.397**	**0.653**	**0.327**
	DL3DV	DINO [31]	15.44	0.443	0.423	0.274	7.49	0.563	0.667	0.408
		VAE [35]	16.37	0.505	0.410	0.294	8.14	0.556	0.643	0.406
		GLD (Ours)	**17.09**	**0.517**	**0.378**	**0.143**	**3.74**	**0.305**	**0.601**	**0.385**

while remaining highly competitive on Re10K. Note that MVGenMaster utilizes an external depth estimator [1] and warps source RGB and depth to the target view as condition inputs, making it heavily reliant on explicit, dense geometric priors. In contrast, GLD operates in a multi-view geometry-aware latent space without requiring explicit geometry conditions (*e.g.*, depth maps) during inference. Although GLD achieves slightly lower pose accuracy on the out-of-domain benchmark than Matrix3D and MVGenMaster, this is expected, as those baselines benefit from strong T2I priors and are additionally fine-tuned on object-centric datasets. Despite relying exclusively on scene-level training without external warping, GLD still achieves the lowest reprojection error and a competitive MEt3R score. This shows that our DA3 latent space effectively encodes the geometric structures necessary for robust, coherent multi-view NVS.

5.3 Qualitative Results

Fig. 4 presents qualitative comparisons of generated target views. The left column (a) shows methods trained from scratch, while the right column (b) shows methods fine-tuned from pretrained T2I models. GLD produces structurally coherent views that closely follow the ground-truth layout, outperforming the VAE and DINOv2 baselines particularly under large viewpoint changes, and remaining competitive with fine-tuned methods despite not leveraging T2I pretraining. We note that the method that relies on warped source views from an external geometry estimator [4] produces sharp outputs but can exhibit noticeable artifacts when the estimation fails, as seen in the fourth column of the first scene, whereas GLD avoids such failure modes. Additional qualitative results are provided in Appendix C.3.

5.4 Decoder Performance

We train the ViT-based RGB decoder, $\mathcal{D}_{\text{rgb}}$, on a combined dataset of Re10K [54] and DL3DV [26]. Following RAE [52], the model is optimized using a weighted sum of $\mathcal{L}_1$, LPIPS, and adversarial losses. We utilize the AdamW optimizer and a cosine learning rate scheduler with a 1-epoch warmup.

Table 5: Geometric Correspondence. PCK results for each DA3 level and DINOv2 [31].

Feature	F_0	F_1	F_2	F_3	DINOv2
PCK ↑	22.25	<u>35.98</u>	**40.70**	20.98	31.64

Table 6: Reconstruction Fidelity. Reconstruction comparison via PSNR and LPIPS across different DA3 [25] levels.

Feature	F_0	F_1	F_2	F_3
PSNR ↑	**28.01**	<u>25.36</u>	14.01	10.19
SSIM ↑	**0.922**	<u>0.873</u>	0.627	0.508
LPIPS ↓	**0.138**	**0.138**	<u>0.491</u>	0.768

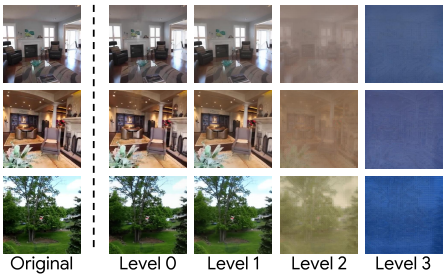


Fig. 5: Reconstruction Visualization. Qualitative comparison showing the reconstruction results for different levels.

The decoder is trained using a multi-resolution strategy involving resolutions of $(504, 504)$, $(504, 378)$, $(504, 336)$, and $(504, 280)$. We employ a global batch size of 128 and an EMA decay of 0.9978. The entire training process is conducted on 8 NVIDIA B200 GPUs, spanning approximately 170k steps. For further comparison, we additionally train a baseline RAE decoder on DINOv2 [31] features using the same architecture and training configuration.

In Tab. 7, we compare $\mathcal{D}_{\text{rgb}}$ with the standard pretrained VAEs used in Stable Diffusion [35] and SDXL [32]. Our decoder reconstructs high-fidelity images from DA3 features and performs competitively with these widely used models. These results indicate that multi-level DA3 features form a sufficiently expressive latent space for high-quality image synthesis.

Table 7: Reconstruction Fidelity Comparison. We evaluate the reconstruction quality using 4,000 randomly sampled images from the Re10K [54] test set.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
VAE [35]	34.53	0.939	0.028
VAE (SDXL) [32]	34.97	0.945	0.029
RAE (DINO) [52]	26.78	0.830	0.148
DA3 Dec. ($\mathcal{D}_{\text{rgb}}$)	**35.41**	**0.960**	**0.019**

5.5 Discussion on boundary layer selection

In § 4.3, we found that boundary layer $k = 1$ yields the best NVS performance. To understand why, we analyze the four levels of features (levels 0 to 3). Specifically, we examine features from geometric and photometric perspectives by measuring their cross-view correspondence and their capacity for high-fidelity image reconstruction, respectively.

For geometric correspondence, we follow the evaluation protocol in Probe3D [10] and assess features on the ScanNet [5] dataset using PCK, the fraction of points that are correctly matched within a distance threshold. Our results in Tab. 5 show that level 1 and level 2 achieve high PCK scores, even outperforming DINOv2 [31]. This indicates that these levels encode stable 3D structures. In contrast, level 0 exhibits poor correspondence; being the shallowest layer, it does not encode the complex geometric relationships needed for cross-view matching.

To evaluate photometric information, we utilize the RGB decoder $\mathcal{D}_{rgb}$ from § 4.2 to reconstruct the original image. Because $\mathcal{D}_{rgb}$ is trained with layer dropout, it is capable of reconstructing images from a single feature level. By reconstructing from each level independently and measuring the error, we find

Table 8: Cascaded vs Independent Generation. We evaluate NVS performance on Re10K with N=4 source views.

Methods	2D Metrics			3D Metrics				
	PSNR ↑	SSIM ↑	LPIPS ↓	ATE ↓	RPE_r ↓	RPE_t ↓	Reproj ↓	MEt3R ↓
Independent	18.81	0.692	0.335	0.197	7.179	0.430	0.666	0.335
Cascaded (Ours)	**19.00**	**0.695**	**0.327**	**0.182**	**6.694**	**0.397**	**0.652**	**0.326**

that levels 0 and 1 retain rich appearance cues, producing accurate colors and textures. However, deeper features, such as level 2, discard essential photometric details. This results in noticeable color loss and smoothed textures (Fig. 5), as well as a notable drop in PSNR compared to shallower levels (Tab. 6).

These results indicate that level 1 successfully meets both requirements. It maintains a high degree of geometric alignment comparable to level 2, while retaining substantially more photometric information than the deeper features. This observation provides a straightforward explanation for why level 1 serves as the optimal latent space for our NVS framework.

5.6 Ablation Studies

Varying the Number of Source Images. In this section, we evaluate the robustness of GLD under varying numbers of source views by comparing against the VAE [35] and DINOv2 [31] baselines. Following the evaluation protocol from our main experiments, we present the quantitative results in Tab. 4. Notably, the margin of improvement in 3D metrics grows as the number of source views decreases. For instance, on DL3DV with N=1, GLD achieves over 3× lower ATE and RPE_r compared to both baselines, whereas with N=4 the gap is approximately 2×. This trend is consistent across both datasets, suggesting that the geometric priors in the DA3 feature space become particularly beneficial when fewer visual cues are available.

Effectiveness of the Cascaded Design. In § 4.4, we generate $\tilde{\mathbf{F}}_0$ by conditioning on the previously synthesized $\tilde{\mathbf{F}}_1$ to ensure consistency across feature levels. To validate this design, we compare against an independent baseline where $\mathcal{M}_0$ and $\mathcal{M}_1$ generate their respective features without seeing each other's output. As shown in Tab. 8, evaluated on Re10K with N=4 source views, the cascading approach consistently improves both 2D and 3D metrics, confirming that allowing shallower features to be conditioned by deeper ones leads to more coherent multi-level representations.

5.7 Geometry Evaluation and 3D Reconstruction

Since GLD generates the multi-level DA3 features, we can leverage the original, pretrained DPT-based decoder $\mathcal{D}_{\text{DA3}}$ to predict ray and depth maps without any additional training or fine-tuning. This allows us to obtain dense geometric predictions as a zero-shot byproduct of the generation process. We evaluate the fidelity of these maps through depth metrics and 3D point-cloud visualizations.

Table 9: Depth Evaluation. Comparison with Matrix3D [29] on ETH3D [37].

Methods	Depth			RGB
	AbsRel↓	SqRel↓	δ_1↑	PSNR↑
Matrix3D [29]	0.197	0.475	0.731	14.13
Ours	**0.160**	**0.410**	**0.800**	**14.80**

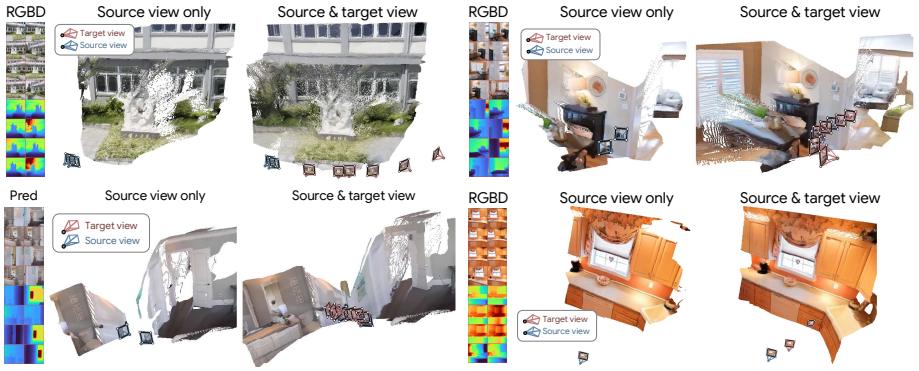


Fig. 6: 3D Reconstruction Visualization. We unproject the synthesized novel views into 3D space using the jointly generated depth and ray maps.

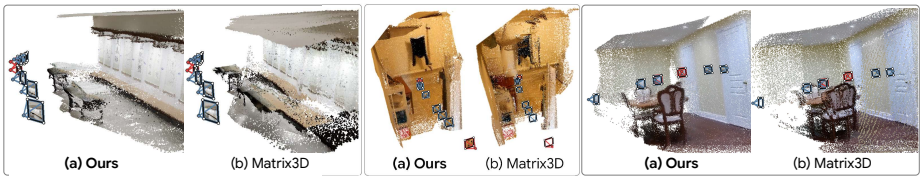


Fig. 7: 3D Reconstruction Comparison. Point clouds unprojected from source and target views RGBD outputs of GLD and Matrix3D [29].

Quantitative Depth Evaluation. We compare our generated depth maps against Matrix3D [29], which jointly generates depth and RGB images. We evaluate performance on the ETH3D dataset [37], which provides high-quality ground-truth depth maps. As shown in Tab. 9, our model produces more accurate depth estimates, suggesting that operating in the DA3 latent space provides effective geometry grounding.

Qualitative 3D Reconstruction Visualization. We visualize the 3D reconstruction by unprojecting the synthesized pixels into 3D space using the generated depth and ray maps. As illustrated in Fig. 6, the resulting point clouds exhibit consistent 3D geometry across diverse camera trajectories, confirming that our generated latent features are geometrically coherent with the underlying scene. We further compare against Matrix3D [29] in Fig. 7. While Matrix3D produces noticeable geometric distortions, GLD yields significantly cleaner and more coherent reconstructions, consistent with the quantitative results in Tab. 9.

6 Conclusion

We presented Geometric Latent Diffusion (GLD), a framework that repurposes the feature space of a geometric foundation model as the latent space for novel view synthesis. Through systematic analysis, we identified a feature level that balances geometric correspondence and photometric fidelity and showed that training diffusion models in this space yields consistent improvements in both

2D image quality and 3D consistency over standard VAE and DINOv2 latent spaces. GLD achieves competitive performance with state-of-the-art methods that leverage large-scale text-to-image pretraining, despite being trained entirely from scratch, while also enabling zero-shot depth and 3D reconstruction as a natural byproduct of the generation process. We hope that this work encourages further investigation into task-specific latent space design for geometry-aware generation.

Acknowledgements

This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, RS-2024-00509279, RS-2025-II212068, RS-2023-00227592, RS-2025-02214479, RS-2024-00457882, RS-2025-25441838, RS-2025-02217259, RS-2026-25519202) and the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025, RS-2024-00333068), and National Research Foundation of Korea (RS-2024-00346597). This research was supported by the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT). This work was completed during WJ and JH's visit to NYU as part of the Global AI Frontier Lab Program. S.X. acknowledges support from Intel Labs, the MSIT IITP grant (RS-2024-00457882) and the NSF award IIS-2443404.

References

1. Depth pro: Sharp monocular metric depth in less than a second. In: International Conference on Learning Representations (2025), https://arxiv.org/abs/2410.02073 11
2. Asim, M., Wewer, C., Wimmer, T., Schiele, B., Lenssen, J.E.: Met3r: Measuring multi-view consistency in generated images. In: IEEE/CVF Computer Vision and Pattern Recognition (CVPR) (2025) 9
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5470–5479 (2022) 7, 8
4. Cao, C., Yu, C., Liu, S., Wang, F., Xue, X., Fu, Y.: Mvgenmaster: Scaling multi-view generation from any image via 3d priors enhanced diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6045–6056 (2025) 2, 3, 8, 9, 11
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017) 12
6. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023) 8
7. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023) 9
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4
9. Du, H., Ye, J., Cong, X., Li, R., Ni, J., Agarwal, A., Zhou, Z., Li, Z., Balestriero, R., Wang, Y.: Videogpa: Distilling geometry priors for 3d-consistent video generation. arXiv preprint arXiv:2601.23286 (2026) 9
10. El Banani, M., Raj, A., Maninis, K.K., Kar, A., Li, Y., Rubinstein, M., Sun, D., Guibas, L., Johnson, J., Jampani, V.: Probing the 3d awareness of visual foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21795–21806 (2024) 12
11. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models. arXiv preprint arXiv:2405.10314 (2024) 2, 3, 6, 8, 9
12. Han, J., Hong, S., Jung, J., Jang, W., An, H., Wang, Q., Kim, S., Feng, C.: Emergent outlier view rejection in visual geometry grounded transformers. arXiv preprint arXiv:2512.04012 (2025) 3
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 2
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) 8
15. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025) 3

16. 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) 3
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 3, 4
18. Kong, X., Watson, D., Strümpler, Y., Niemeyer, M., Tombari, F.: Causnvs: Autoregressive multi-view diffusion for flexible 3d novel view synthesis. arXiv preprint arXiv:2509.06579 (2025) 3
19. Krishnan, A., Yan, X., Casser, V., Kundu, A.: Orchid: Image latent diffusion for joint appearance and geometry generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28217–28227 (2025) 3
20. Kwak, M.S., Kim, J., Yun, S., Han, D., Kim, T., Kim, S., Kim, J.H.: Aligned novel view image and geometry synthesis via cross-modal attention instillation. arXiv preprint arXiv:2506.11924 (2025) 2, 3
21. Kwon, M., Choi, J., Park, J., Jeon, S., Jang, J., Seo, J., Kwak, M., Kim, J.H., Kim, S.: Cameo: Correspondence-attention alignment for multi-view diffusion models. arXiv preprint arXiv:2512.03045 (2025) 3, 8, 9
22. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024) 2
23. Li, L., Zhang, Z., Li, Y., Xu, J., Hu, W., Li, X., Cheng, W., Gu, J., Xue, T., Shan, Y.: Nvcomposer: Boosting generative novel view synthesis with multiple sparse and unposed images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 777–787 (2025) 3, 8, 9
24. Li, R., Yi, B., Liu, J., Gao, H., Ma, Y., Kanazawa, A.: Cameras as relative positional encoding. In: Advances in Neural Information Processing Systems (2025) 6
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) 1, 2, 3, 4, 5, 8, 12
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) 7, 8, 11
27. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023) 6
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 8
29. Lu, Y., Zhang, J., Fang, T., Nahmias, J.D., Tsin, Y., Quan, L., Cao, X., Yao, Y., Li, S.: Matrix3d: Large photogrammetry model all-in-one. Computer Vision and Pattern Recognition (CVPR) (2025) 3, 8, 9, 13, 14
30. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 3
31. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 3, 4, 8, 11, 12, 13
32. 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) 2, 12

33. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10901–10911 (2021) 9
34. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: International Conference on Computer Vision (ICCV) 2021 (2021) 7
35. 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) 1, 2, 3, 4, 8, 11, 12, 13
36. 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) 3
37. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 13, 14
38. 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) 2, 3
39. Shi, M., Wang, H., Zheng, W., Yuan, Z., Wu, X., Wang, X., Wan, P., Zhou, J., Lu, J.: Latent diffusion model without variational autoencoder. arXiv preprint arXiv:2510.15301 (2025) 2, 3
40. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023) 2
41. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: Dinov3 (2025), https://arxiv.org/abs/2508.10104 3
42. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020) 2
43. Tong, S., Zheng, B., Wang, Z., Tang, B., Ma, N., Brown, E., Yang, J., Fergus, R., LeCun, Y., Xie, S.: Scaling text-to-image diffusion transformers with representation autoencoders. arXiv preprint arXiv:2601.16208 (2026) 2, 3
44. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025) 3, 4
45. 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) 1, 2, 3, 4, 5, 8, 9
46. Wang, S., Tian, Z., Huang, W., Wang, L.: Ddt: Decoupled diffusion transformer. arXiv preprint arXiv:2504.05741 (2025) 6
47. 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) 3

48. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam (2020) 7
49. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025) 2, 3, 4, 5
50. Yang, Y., Shao, J., Li, X., Shen, Y., Geiger, A., Liao, Y.: Prometheus: 3d-aware latent diffusion models for feed-forward text-to-3d scene generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2857–2869 (2025) 3
51. 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. arXiv preprint arXiv:2409.02048 (2024) 2, 3
52. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025) 1, 2, 3, 6, 11, 12
53. 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) 3
54. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. ACM Trans. Graph. (Proc. SIGGRAPH) **37** (2018), https://arxiv.org/abs/1805.09817 6, 7, 8, 11, 12