Unlocking 3D in Video DiTs: Camera Adapters with SfM Supervision

Anonymous CVPR submission

Paper ID *****

Abstract

Video diffusion transformers (DiTs), trained on large-scale video data, have been shown to encode rich 3D spatial priors. We show that this latent understanding can be made explicit for novel view synthesis through two lightweight interventions: (1) parallel camera attention adapters *that inject relative camera geometry directly into attention through camera-aware positional encoding. (2) a* sparse correspondence loss *that uses sparse SIFT correspondences verified by known-camera epipolar geometry to directly supervise adapter query-key projections with an InfoNCE objective. Adding only* ***2.7%*** *trainable parameters to a frozen 1.3B-parameter video DiT, our method generates a variable number of geometrically consistent novel views (up to 20) at arbitrary camera angles from one or more input images in a single end-to-end pass. On the SeVA benchmark, we outperform all prior methods on two of three evaluation splits, achieving up to 28.1 dB PSNR, while being significantly more parameter-efficient than dedicated NVS models. We further show that generated views are geometrically coherent enough for downstream 3D Gaussian Splatting reconstruction, producing 3D scene representations from a single photograph.*

1. Introduction

Open-world 3D scene understanding [8, 16, 18] has made remarkable progress, yet these methods fundamentally rely on multi-view captures or RGB-D scans as input. Obtaining such data at scale remains a bottleneck: dense captures require physical access and specialized sensors, limiting applicability to scenes that can be actively scanned. Novel view synthesis (NVS) offers a compelling alternative; generating geometrically consistent multi-view imagery from as little as a single photograph, potentially enabling 3D understanding of any scene for which at least one image exists. Modern video diffusion transformers (DiTs) [20, 27] trained on internet-scale video data have been characterized as "world models" [1] that implicitly learn rich priors about 3D structure, object permanence, and scene dynamics. A natural question arises: *can we unlock these latent 3D priors for explicit view synthesis, without training a dedicated NVS model from scratch?*

Existing diffusion-based NVS methods [4, 14, 29, 32] train or fine-tune entire architectures for the task, requiring substantial compute and large-scale multi-view datasets. For example, Stable Virtual Camera (SeVA) [32] fine-tunes a full Stable Video Diffusion model and requires a per-scene oracle camera scale sweep at inference.

We propose a significantly more parameter-efficient alternative that keeps the video DiT backbone *entirely frozen*, injecting camera-awareness through lightweight parallel attention adapters and reinforcing geometric consistency with a correspondence loss derived from known camera geometry. Our approach is built on two key components:

(1) Parallel Camera Attention Adapters. We introduce small parallel attention branches into each DiT block that process the same hidden features as the frozen self-attention but with *Projective Rotary Position Encoding* (PRoPE) [13] applied to queries and keys. PRoPE encodes camera geometry via projection-matrix-derived rotary embeddings, so that dot-product attention naturally becomes a function of relative camera pose. The adapter outputs are added to the frozen attention outputs via a zero-initialized output projection, preserving pretrained generation quality at initialization.

(2) SfM Correspondence Supervision. Beyond the standard flow matching loss, we apply an InfoNCE [15] contrastive loss on the adapter's PRoPE-transformed query-key pairs using sparse 2D correspondences obtained via SIFT matching and epipolar filtering with known camera poses. This explicitly trains the adapter to match features across views at geometrically corresponding locations, providing a direct 3D-aware training signal.

Our method adds only 35.4M trainable parameters (**2.7%** of 1.3B) and generates a variable number of novel views (up to 20) at *arbitrary camera angles* from *one or more* conditioning images in a single end-to-end pass. We train on the RealEstate10K dataset [33] (∼54K scenes) and evaluate on the SeVA benchmark [32] across three standard splits. Critically, we show that generated views are geo-

metrically coherent enough to fit 3D Gaussian Splatting [9] models, producing explicit 3D scene representations from a single input image, a pipeline that could directly feed downstream 3D scene understanding.

Our contributions are:

- Lightweight **camera attention adapters** with PRoPE that unlock NVS in a frozen video DiT at arbitrary camera angles, adding only 2.7% trainable parameters.
- **Geometric correspondence supervision** via InfoNCE on PRoPE-transformed adapter Q,K pairs, providing explicit geometric training signal from sparse SIFT correspondences verified by epipolar geometry.
- Evaluation on the **SeVA benchmark** with ablations and **downstream 3DGS reconstruction**, demonstrating a single-image-to-3D pipeline.

2. Related Work

Novel View Synthesis with Diffusion Models. Diffusion-based NVS has progressed from single-view methods [14, 17] to multi-view generators that jointly produce consistent views. CAT3D [4] scales multi-view image diffusion for fast 3D reconstruction, while EscherNet [11] introduces a camera positional encoding for continuous relative view control, and EpiDiff [7] constrains cross-view attention to epipolar lines for geometric consistency. A growing line of work repurposes *video* diffusion for NVS by reinterpreting the temporal axis as viewpoint change: SV3D [19] adapts SVD for orbital multi-view synthesis but requires full fine-tuning and is limited to object-centric data, ViewCrafter [29] conditions video diffusion on point-cloud-warped frames, CamCo [25] adds camera control via Plücker coordinates with epipolar attention, and SeVA [32] fine-tunes the entire SVD backbone end-to-end. Unlike these approaches that either train full models or rely on UNet backbones, we adapt a frozen video DiT with minimal parameters, treating the pretrained model as a world model whose 3D priors are unlocked through lightweight camera injection.

Camera Control and Adaptation in Generative Models. Camera control in diffusion models ranges from learned trajectory embeddings [23] to geometric encodings such as PRoPE [13], which encodes full camera frustums as a relative positional encoding in multi-view transformers. On the adaptation side, ControlNet [31] pioneered frozen-backbone adaptation through parallel encoder branches, and IP-Adapter [28] adds image conditioning via decoupled cross-attention. Our camera attention adapters combine both lines of work: PRoPE-based geometric encoding within lightweight parallel attention heads alongside the frozen self-attention.

Geometric Supervision in Generation. Explicit geometric losses in diffusion training remain underexplored. CAMEO [12] supervises attention maps with geometric correspondence in UNet-based multi-view diffusion models. We similarly leverage correspondence supervision, but apply it to PRoPE-transformed adapter Q,K pairs via InfoNCE [15] rather than direct attention map supervision.

Feed-Forward 3D Reconstruction and Scene Understanding. An alternative to generative NVS is feed-forward reconstruction: DUSt3R [22] and VGGT [21] predict dense 3D point maps from arbitrary image sets, while pixelSplat [2], MVSplat [3], DepthSplat [26], and GS-LRM [30] predict 3D Gaussian primitives from sparse posed views. These methods excel on narrow-baseline interpolation but struggle with large viewpoint extrapolation where synthesis of unseen content is required. Our generative approach is complementary: pretrained video priors enable plausible synthesis at distant viewpoints, while the resulting views can serve as input for downstream reconstruction (Sec. 4.5) and 3D scene understanding [8, 10, 16, 18].

3. Method

Given one or more input images $\{I_p\}_{p=1}^{P}$ and a set of target camera poses $\{\mathbf{T}_n, \mathbf{K}_n\}_{n=1}^{N}$ (extrinsics and intrinsics), our goal is to generate N geometrically consistent novel views in a single end-to-end pass through a pretrained video DiT (see Fig. 1 for an overview). Both P and N are flexible, and the target cameras can specify arbitrary viewpoints without restriction to fixed grids or predefined trajectories.

3.1. Backbone

We adopt Wan 2.1 (1.3B parameters) [20] as our backbone, a flow matching video DiT whose image-to-video variant conditions generation on a single frame. The model produces 81-frame videos at 576×576 resolution through a 3D VAE with $4\times$ temporal compression, yielding 21 latent frames.

View-to-latent packing. Since the 3D VAE compresses every 4 consecutive pixel-space frames into one latent frame, we replicate each of the N target views 4 times while the conditioning view occupies a single frame ($1 + N \times 4 = 81$ pixel frames $\rightarrow$ 21 latent frames, giving $P + N \leq 21$). This yields a 1:1 mapping between views and latent frames, allowing us to assign a unique camera pose to each temporal token.

3.2. Camera Attention Adapters

The core of our approach is a set of parallel camera attention branches injected into each of the 30 DiT blocks. The $8\times$ VAE spatial compression followed by the DiT's $2\times$ patch embedding reduce each 576×576 frame to a 36×36 token

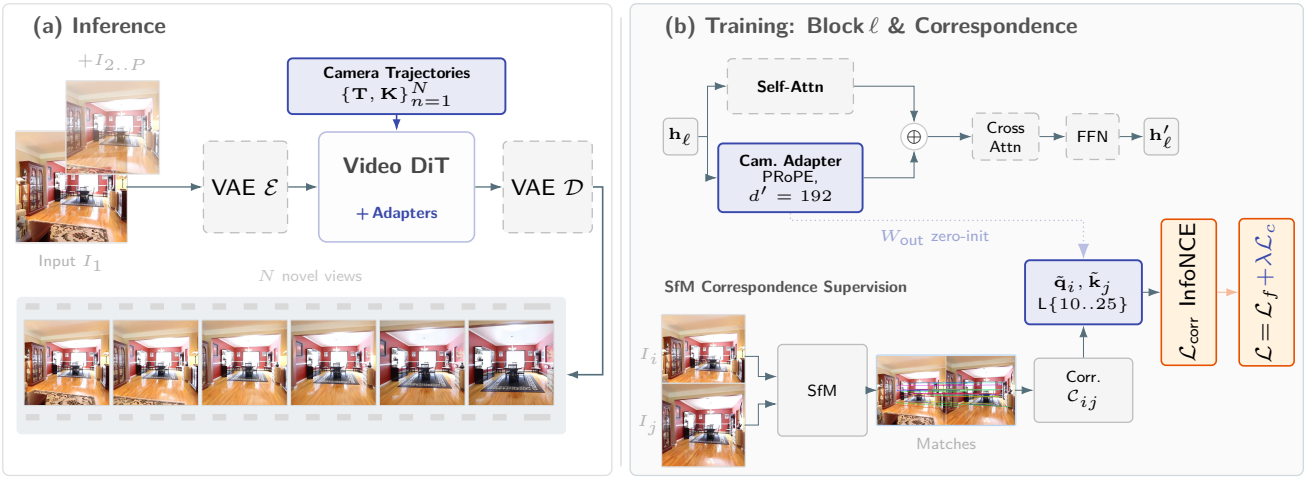

Figure 1. **Overview of CamAdapter. (a) Inference:** Given one or more input images $\{I_p\}$ and target camera poses $\{\mathbf{T}, \mathbf{K}\}_{n=1}^{N}$, we generate N novel views through a frozen video DiT augmented with lightweight camera adapters (35.4M trainable parameters, 2.7% of 1.3B). **(b) Training:** Each DiT block ℓ contains a frozen self-attention and a parallel camera adapter that injects PRoPE-encoded camera poses ($d' = 192$). For geometric correspondence supervision, sparse SIFT correspondences verified by epipolar geometry from known cameras supervise the adapter's query-key features $(\tilde{\mathbf{q}}_i, \tilde{\mathbf{k}}_j)$ at layers $\{10, 15, 20, 25\}$ through an InfoNCE loss. The total loss combines the flow matching objective $\mathcal{L}_f$ with the correspondence loss $\lambda \mathcal{L}_c$.

grid. All frames and spatial positions are flattened into a single sequence of length $S = F \times H' \times W'$ ($F = 21$, $H' = W' = 36$), so each block ℓ operates on features $\mathbf{h}_\ell \in \mathbb{R}^{B \times S \times d}$ with $d = 1536$. The standard frozen self-attention computes:

$$\mathbf{h}_\ell^{\text{self}} = \text{SelfAttn}_\ell(\mathbf{h}_\ell) \tag{1}$$

Our camera adapter operates in parallel with a compressed hidden dimension $d' = d/c = 192$ (compression ratio $c = 8$):

$$\mathbf{Q} = \mathbf{h}_\ell \cdot W_Q, \quad \mathbf{K} = \mathbf{h}_\ell \cdot W_K, \quad \mathbf{V} = \mathbf{h}_\ell \cdot W_V \tag{2}$$

$$\tilde{\mathbf{Q}} = \text{PRoPE}_Q(\mathbf{Q}), \quad \tilde{\mathbf{K}} = \text{PRoPE}_K(\mathbf{K}) \tag{3}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d'}$ project to the compressed dimension. The PRoPE encoding [13] applies *complementary* camera-derived transforms to queries and keys: PRoPE_Q applies $\Pi^\top$ (the transpose of the camera projection matrix) while PRoPE_K applies Π^{-1} (its inverse), each combined with 2D spatial rotary embeddings. This causes the dot-product $\tilde{\mathbf{Q}}^\top \tilde{\mathbf{K}}$ to become a function of the *relative* camera geometry between any two views.

The adapter output is projected back to the original dimension through a zero-initialized output projection:

$$\mathbf{h}'_\ell = \mathbf{h}_\ell^{\text{self}} + W_{\text{out}} \cdot \text{Attn}(\tilde{\mathbf{Q}}, \tilde{\mathbf{K}}, \mathbf{V}) \tag{4}$$

with $W_{\text{out}} \in \mathbb{R}^{d' \times d}$ initialized to zero, so the adapter has no effect at initialization and the pretrained model is fully preserved. With 2 attention heads and 192-dimensional projections per block, the total adapter overhead is only ∼35.4M parameters (2.7% of the 1.3B backbone).

3.3. Multi-Frame Conditioning

During training, we condition on a variable number of ground-truth views: a primary input view I_1 plus $k \sim \text{Uniform}(0, K)$ additional views ($K = 5$). The conditioning views are encoded by the VAE and injected via the input conditioning mechanism of the image-to-video DiT, with binary masks indicating which latent frames are conditioned. This enables flexible conditioning at inference: the model can take a single photograph and generate up to 20 novel views, or leverage multiple input views (e.g., $P = 3$) for improved geometric consistency.

3.4. SfM Correspondence Supervision

Standard flow matching training provides a per-token reconstruction signal in latent space but does not explicitly enforce geometric consistency across views. To address this (Fig. 2), we leverage the known camera poses and intrinsics available in the training data to extract sparse 2D-2D correspondences offline. For each pair of training frames, we detect SIFT keypoints and match them using a ratio test. We then compute the fundamental matrix from the known cameras and discard any match whose symmetric epipolar distance exceeds a threshold, retaining only geometrically verified correspondences. At training time, these correspondences are mapped to the 36×36 latent token grid and used to supervise the adapter's internal query–key representations.

Contrastive formulation. A naive approach would directly supervise the full attention matrix at corresponding

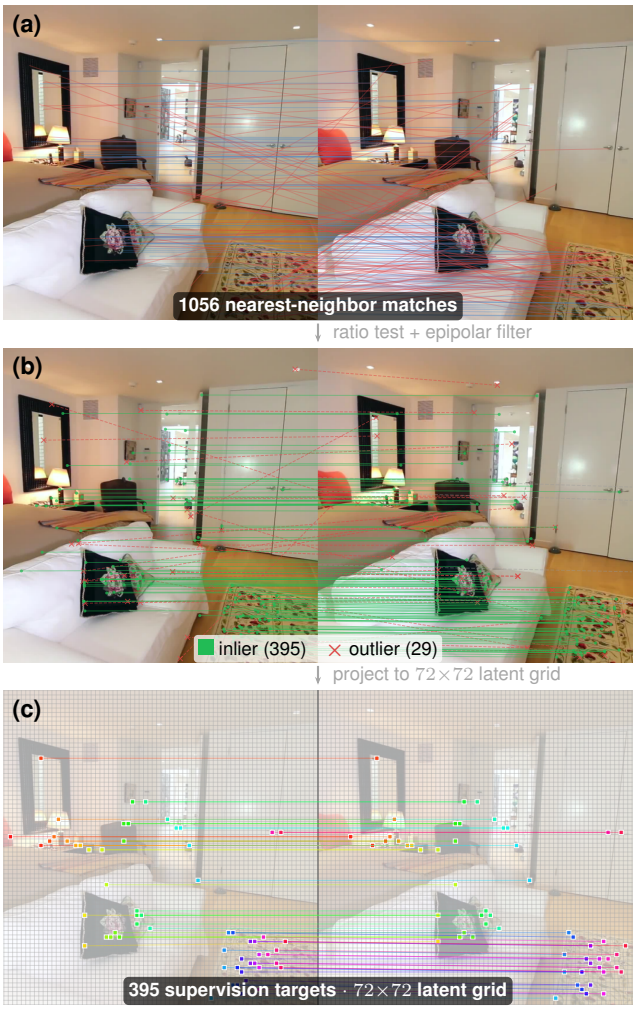

Figure 2. **Correspondence supervision pipeline.** **(a)** SIFT nearest-neighbor matching between a training frame pair yields many noisy correspondences. **(b)** Lowe's ratio test and epipolar filtering using the known cameras retains only geometrically consistent matches (green), rejecting outliers (red). **(c)** Verified correspondences are projected to the 72×72 latent grid, where they provide supervision targets for the adapter's query-key features via InfoNCE.

positions; however, with $S > 27$K tokens, materializing $\mathbf{A}^{(\ell)} \in \mathbb{R}^{S\times S}$ at multiple layers is memory-prohibitive. We instead formulate correspondence supervision as a sparse InfoNCE objective computed directly from the adapter's query and key projections, requiring only $O(M \times N_{\text{neg}})$ operations for M correspondence pairs rather than $O(S^2)$.

For a pair of corresponding spatial tokens (i, j) across views, we extract the adapter's PRoPE-transformed query $\tilde{\mathbf{q}}_i$ and key $\tilde{\mathbf{k}}_j$ at layers $\ell \in \{10, 15, 20, 25\}$, corresponding to middle-to-deep blocks where geometric reasoning is most pronounced, and apply:

$$\mathcal{L}_{\text{corr}} = -\log \frac{\exp(\tilde{\mathbf{q}}_i^\top \tilde{\mathbf{k}}_j/\tau)}{\sum_{j'} \exp(\tilde{\mathbf{q}}_i^\top \tilde{\mathbf{k}}_{j'}/\tau)} \tag{5}$$

where $\tau = 0.5$ is the temperature and negatives are 128 randomly sampled non-corresponding spatial positions within the same latent feature maps. This loss directly trains the adapter to produce camera-aware features that match across views at geometrically corresponding locations.

To prevent the correspondence signal from interfering with early flow matching convergence, we apply $\mathcal{L}_{\text{corr}}$ with a warmup curriculum that linearly ramps λ from 0 to its target value. The total training loss is:

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda \cdot \mathcal{L}_{\text{corr}} \tag{6}$$

4. Experiments

4.1. Setup

Training. We train on the RealEstate10K dataset [33], comprising ∼54K indoor and outdoor video sequences with camera poses. Each scene contains 30 extracted frames; at each iteration we randomly sample 21 of them, providing view diversity across epochs. The first frame serves as the primary conditioning view, while the remaining 20 are generation targets (with up to $K = 5$ additional frames randomly promoted to conditioning views, as described in Sec. 3.2). We train for 20 epochs on 8×A40 GPUs using AdamW with a learning rate of 1×10^{-4} for the adapters. The correspondence loss is disabled for the first 600 steps, then λ is linearly ramped from 0 to 0.01 over the next 600 steps.

Evaluation. We adopt the benchmark protocol from SeVA [32], which defines three evaluation splits: **R** (10 RE10K scenes, 11 views, evaluating scene reconstruction quality), **V** (10 ViewCrafter scenes, 25 views, evaluating long-range view consistency), and **D** (128 RE10K scenes, 7 views per scene, evaluating diversity across scenes). We report PSNR, SSIM, and LPIPS following the exact SeVA protocol. For fair comparison, we report SeVA results at cfg = 2.0 (the paper's default) with per-scene oracle camera scale selection over 20 scales.

Baselines. Following the SeVA [32] benchmark, we compare against all methods reporting results on the RE10K splits. These include *regression-based* models: MVSplat [3] and DepthSplat [26] (feed-forward Gaussian splatting), as well as *diffusion-based* models: MotionCtrl [23], 4DiM [24], ViewCrafter [29], and SeVA [32].

Table 1. **PSNR↑ on the SeVA benchmark.** P denotes the number of input views. For SeVA with $P=1$, per-scene oracle scale sweep over 20 scales is applied. Time is per-scene on the R split ($P=1$) for diffusion-based methods; †SeVA total with oracle sweep is ∼90 min/scene.

Method	Type	R split		V split	D split	Time
		P=1	P=3	P=1	P=1	(s)
Regression-based						
MVSplat [3]	Feed-fwd	21.56	25.64	20.32	20.42	—
DepthSplat [26]	Feed-fwd	21.87	22.54	19.24	20.90	—
Diffusion-based						
MotionCtrl [23]	Full ft	—	—	16.29	12.74	∼80
4DiM [24]	Full ft	—	—	—	17.08	—
ViewCrafter [29]	Full ft	20.88	22.81	22.04	20.43	∼120
SeVA [32]	Full ft	18.11	27.57	18.56	17.99	∼270†
Ours	Adapter	21.24	**28.13**	**22.49**	19.67	**∼67**

4.2. Main Results

Table 1 compares our method against all baselines from the SeVA benchmark on RE10K. Our method achieves the best results on two of three evaluation splits, despite training only 2.7% of the backbone parameters.

On the R split with multi-view input ($P=3$), we achieve 28.13 dB, surpassing SeVA (27.57 dB) by 0.56 dB and the best regression-based model MVSplat (25.64 dB) by 2.49 dB. On the V split ($P=1$), which evaluates long-range view consistency across 25 views, we achieve 22.49 dB, outperforming ViewCrafter (22.04 dB) by 0.45 dB and substantially exceeding DepthSplat (19.24 dB) by 3.25 dB, highlighting the advantage of generative models for large camera motions where feed-forward splatting methods produce blurry renders. On the R split with $P=1$, we achieve 21.24 dB, competitive with the regression-based DepthSplat (21.87 dB) while being the best among all diffusion-based methods.

Comparison with regression-based methods. MVSplat and DepthSplat achieve strong PSNR on the R and D splits where the evaluation views are close to the input cameras. However, on the V split, which requires rendering 25 views along a long camera trajectory, both methods degrade significantly (MVSplat: 20.32 dB, DepthSplat: 19.24 dB vs. ours: 22.49 dB), as feed-forward splatting struggles with large viewpoint extrapolation beyond the input pair.

Comparison with diffusion-based methods. Among generative approaches, our method achieves the best quality while being the fastest: ∼67s vs. ∼120s for ViewCrafter and ∼270s for a single SeVA pass (which requires repeating 20× for oracle scale selection at $P=1$). Notably, SeVA fine-tunes the *entire* SVD backbone end-to-end, whereas we train only lightweight adapters (2.7% of parameters) on top of a *frozen* video DiT.

4.3. Qualitative Results

Figures 3 and 4 show qualitative comparisons on six representative scenes spanning both the RE10K and ViewCrafter evaluation protocols. On Scene 1 ($P=3$), both methods produce plausible views, but our approach better preserves the wall artwork and room geometry as the camera moves through the interior. On the single-view RE10K scene (Scene 2), our method faithfully reproduces the corridor's receding perspective and the chair position, whereas SeVA fail to follow the camera trajectory faithfully. On the ViewCrafter scene (Scene 3), which involves a trajectory through an interior with distinctive wooden doors and carpet, SeVA shows noticeable color drift, while our method maintains stable appearance and geometry across the full trajectory and go all the way in the corridor. These observations are consistent with the quantitative gains reported in Table 1.

Figure 4 presents additional qualitative comparisons on three further scenes. On Scene 4 ($P=3$), our method accurately reconstructs the bedroom layout and furnishings throughout the trajectory, whereas SeVA catastrophically degrades into saturated orange artifacts in the latter views. On Scene 5 ($P=1$, RE10K), our generated views maintain crisp room geometry and consistent lighting, while SeVA exhibits noticeable color shifts and blurring. On Scene 6 ($P=1$, ViewCrafter), involving a long-range trajectory through a hallway with framed artwork, our method preserves scene appearance and structural details while SeVA shows progressive degradation.

4.4. Ablation Study

Table 2 ablates the components of our method across all evaluation splits.

Camera adapters are the dominant component: replacing our adapters and correspondence loss with LoRA only ("LoRA only") causes a dramatic degradation across all splits (−3.03 dB on $R_{P=1}$, −8.09 dB on $R_{P=3}$, −2.30 dB on $V_{P=1}$), confirming that PRoPE-based geometric injection is essential and that LoRA alone cannot learn camera control from the training data.

Correspondence loss provides consistent improvement: removing the InfoNCE supervision while keeping adapters causes PSNR to drop by 0.67 dB on $R_{P=1}$ and 0.35 dB on $R_{P=3}$, while LPIPS consistently worsens, demonstrating that explicit geometric guidance complements PRoPE-based geometric injection.

LoRA does not improve results when combined with adapters: adding rank-16 LoRA [6] on top of the full model ("w/ LoRA") slightly degrades performance on $R_{P=1}$ (−0.23 dB) and $R_{P=3}$ (−0.06 dB). On $V_{P=1}$, LoRA yields

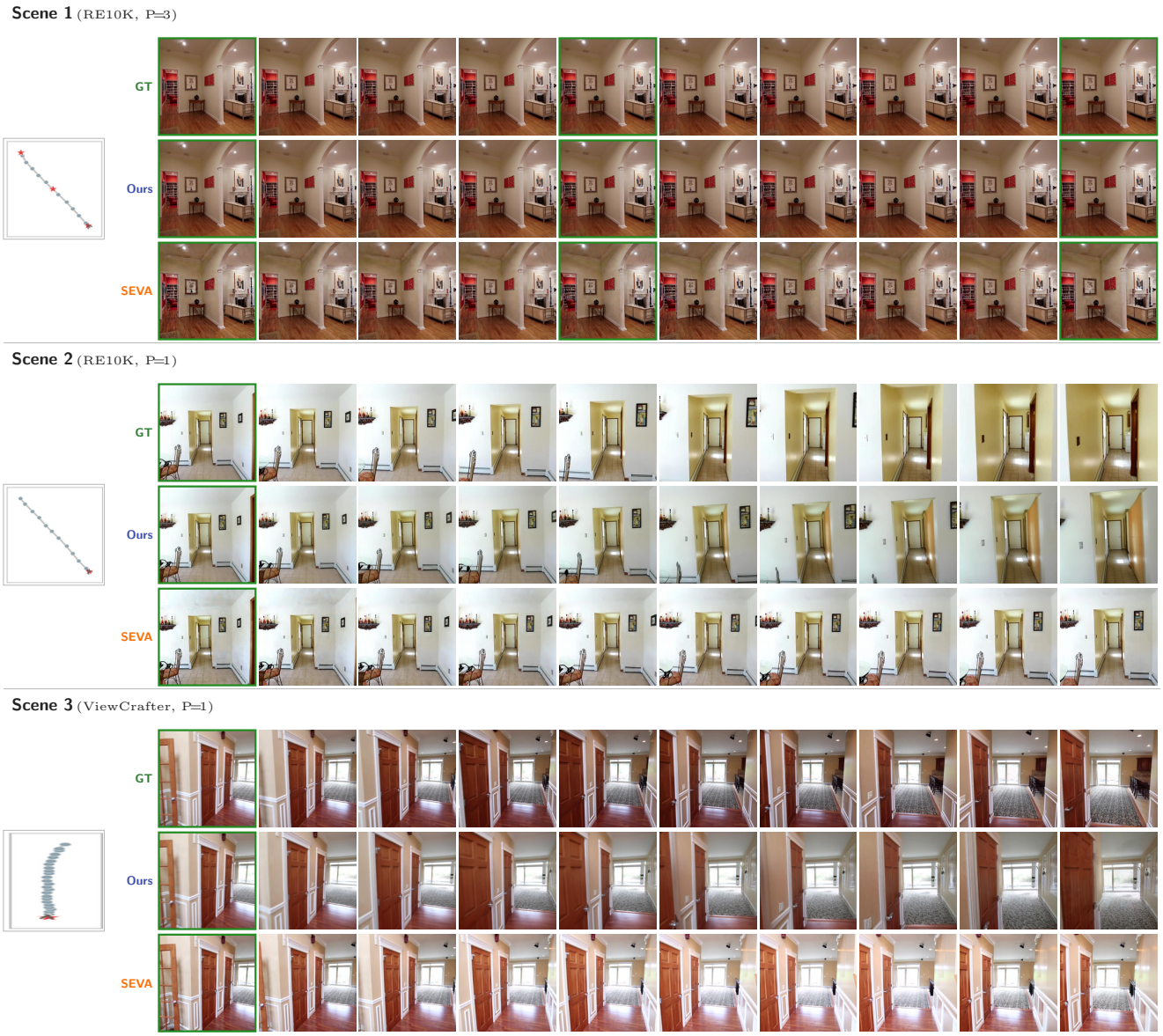

Figure 3. **Qualitative comparison on three scenes.** Each block shows the camera trajectory (left) and 10 sampled views along it for SeVA, CAMADAPTER (Ours), and ground truth (GT). Green borders indicate input (conditioning) views. Scenes 1–2 are from the RE10K R split with three ($P=3$) and one ($P=1$) input views respectively; Scene 3 uses a single input ($P=1$) from the ViewCrafter V split. Our method produces views that are visually consistent with GT across the trajectory, while SeVA exhibits noticeable color shifts and geometric distortions, particularly in the single-view scenarios.

Table 2. **Ablation study across evaluation splits.** Each row varies the component composition. The full model uses camera adapters with correspondence loss and no LoRA.

	$R_{P=1}$			$R_{P=3}$			$V_{P=1}$		
Configuration	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Full model	**21.24**	**.752**	**.248**	**28.13**	**.892**	**.058**	22.49	**.768**	**.232**
w/o corr. loss	20.57	.726	.277	27.78	.882	.059	22.22	.747	.243
LoRA only	18.21	.644	.410	20.04	.687	.238	20.19	.701	.318
w/ LoRA	21.01	.738	.272	28.07	.890	.059	**22.60**	.760	.237

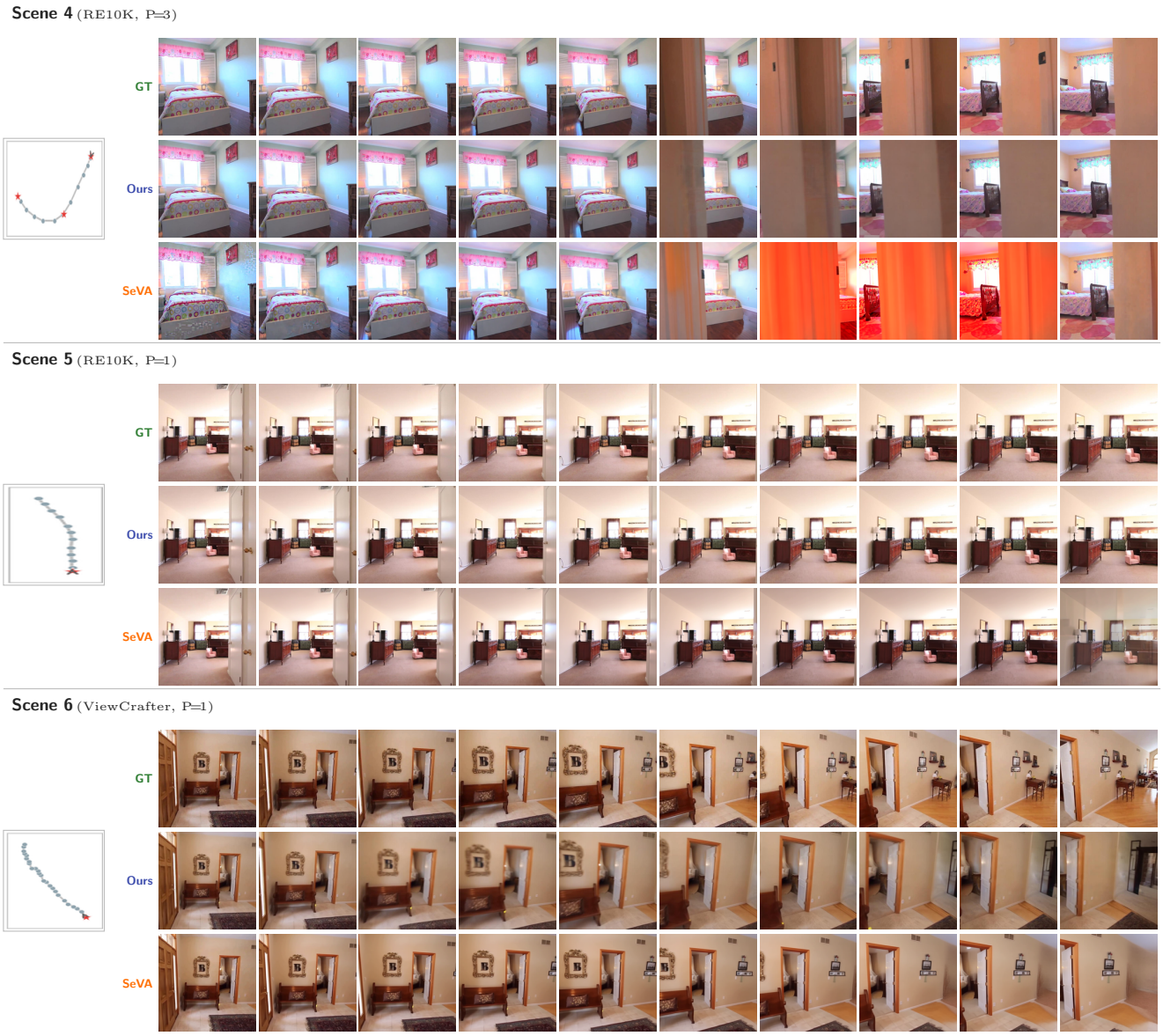

Figure 4. **Additional qualitative comparison.** Three further scenes: an RE10K bedroom with $P = 3$ input views (Scene 4), an RE10K living room with $P = 1$ (Scene 5), and a ViewCrafter hallway with $P = 1$ (Scene 6). Across all scenes, CAMADAPTER produces sharper, more geometrically faithful novel views than SeVA, which suffers from severe color drift and structural collapse, particularly under large viewpoint changes.

a marginal PSNR gain (+0.11 dB) but worse SSIM and LPIPS, suggesting that the additional representational capacity is unnecessary given the frozen backbone's strong priors. We therefore use the model without LoRA as our final configuration.

Figure 5 provides qualitative support for these findings on an example RE10K scene. Near the input view, the full model and w/o correspondence loss variants produce nearly indistinguishable results; however, at distant views the full model better preserves fine details such as the sofa edges and the chair position, reflecting the geometric guidance provided by InfoNCE supervision. Adding LoRA on top of adapters (w/ LoRA) yields comparable visual quality to the full model, consistent with the marginal quantitative differences in Table 2. The LoRA-only variant, which lacks camera adapters entirely, fails catastrophically: generated views has no resemblance to the target scene and instead produce incoherent content, underscoring that explicit camera conditioning through PRoPE is indispensable and cannot be compensated by LoRA's representational capacity alone.

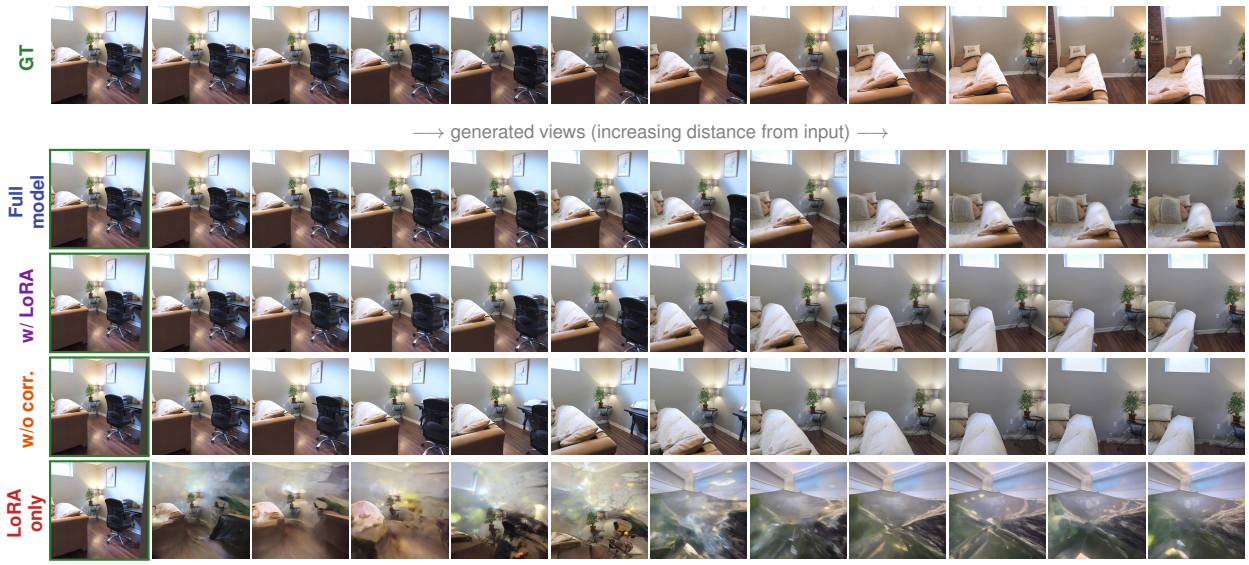

Figure 5. **Qualitative ablation** on an RE10K scene ($P = 1$). All 11 generated views are shown per row, ordered by increasing distance from the single input view (green border). **Full model** and **w/o corr. loss** produce comparable quality near the input but the full model better preserves furniture geometry at distant views. **w/ LoRA** performs similarly to the full model. **LoRA only** (no camera adapters) completely fails to generate the target scene.

Table 3. **Downstream 3DGS reconstruction.** We fit 3DGS to generated views and render held-out ground-truth cameras. CAMADAPTER produces more 3D-consistent views, yielding uniformly better reconstruction quality across all splits and metrics.

Split	Method	PSNR↑	SSIM↑	LPIPS↓
$R_{P=1}$	SeVA	13.35	0.468	0.523
	CAMADAPTER	**13.84**	**0.474**	**0.514**
$R_{P=3}$	SeVA	19.04	0.672	0.329
	CAMADAPTER	**19.39**	**0.685**	**0.293**
$V_{P=1}$	SeVA	11.32	0.395	0.619
	CAMADAPTER	**14.98**	**0.512**	**0.469**
Overall	SeVA	14.45	0.508	0.495
	CAMADAPTER	**16.03**	**0.555**	**0.427**

4.5. Downstream 3D Reconstruction

A key motivation for geometrically consistent NVS is enabling 3D scene representations from minimal input, bridging the gap between single-image observation and the multi-view data required by 3D scene understanding methods [8, 16]. To evaluate this, we fit 3D Gaussian Splatting (3DGS) [9] models to our generated views using their associated camera poses, then render novel viewpoints not seen during fitting.

We run this experiment across all benchmark scenes and splits, fitting 3DGS to the generated views plus input frames and evaluating rendered quality on up to 30 held-out ground-truth viewpoints per scene. Table 3 reports per-split averages. CAMADAPTER outperforms SeVA across all splits and all metrics, with the largest gains on ViewCrafter (+3.66 dB PSNR, −0.15 LPIPS), confirming that more 3D-consistent generated views translate directly into better downstream reconstruction quality.

This experiment validates that our generated views are not only visually plausible but carry sufficient geometric coherence for explicit 3D reconstruction across diverse scenes. The resulting 3D representations could serve as input for downstream tasks such as open-vocabulary 3D segmentation [18], language-grounded scene understanding [16], or 3D scene graphs [5], extending 3D understanding to any scene captured in a single photograph.

5. Conclusion

We have shown that pretrained video diffusion transformers can be effectively adapted for novel view synthesis through lightweight camera attention adapters (PRoPE) and geometric correspondence supervision (InfoNCE), adding only 2.7% trainable parameters. Our approach generates geometrically consistent multi-view images in a single end-to-end pass, and the resulting views support downstream 3D Gaussian Splatting reconstruction — enabling a single-image-to-3D pipeline. This suggests that large video DiTs already possess substantial 3D spatial understanding that can be unlocked with minimal, targeted intervention, offering a practical and parameter-efficient path toward generating the geometrically consistent multi-view data that open-world 3D scene understanding methods critically require.

References

[1] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024. 1

[2] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19457–19467, 2024. 2

[3] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European conference on computer vision*, pages 370–386. Springer, 2024. 2, 4, 5

[4] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 1, 2

[5] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024. 8

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5

[7] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9784–9794, 2024. 2

[8] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023. 1, 2, 8

[9] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 8

[10] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19729–19739, 2023. 2

[11] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9503–9513, 2024. 2

[12] Minkyung Kwon, Jinhyeok Choi, Jiho Park, Seonghu Jeon, Jinhyuk Jang, Junyoung Seo, Minseop Kwak, Jin-Hwa Kim, and Seungryong Kim. Cameo: Correspondence-attention alignment for multi-view diffusion models. *arXiv preprint arXiv:2512.03045*, 2025. 2

[13] Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *arXiv preprint arXiv:2507.10496*, 2025. 1, 2, 3

[14] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 1, 2

[15] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 2

[16] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 815–824, 2023. 1, 2, 8

[17] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9420–9429, 2024. 2

[18] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023. 1, 2, 8

[19] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2024. 2

[20] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2

[21] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2

[22] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20697–20709, 2024. 2

[23] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 4, 5

[24] Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J Fleet. Controlling space and time with diffusion models. *arXiv preprint arXiv:2407.07860*, 2024. 4, 5

[25] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 2

[26] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16453–16463, 2025. 2, 4, 5

[27] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1

[28] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2

[29] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 1, 2, 4, 5

[30] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 2

[31] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2

[32] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025. 1, 2, 4, 5

[33] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 1, 4