ExPose: Reinforcing Video Generation Models for Extreme Pose Estimation

Youngho Yoon[1*], Wonjune Cho[2], Hyunho Ha[2], Sujung Kim[2], and Kuk-Jin Yoon[1]
[1]Visual Intelligence Lab., KAIST [2]NAVER LABS

Abstract

*Pose estimation remains challenging under sparse views, especially when visual overlap across images is extremely limited. Recent advances in video generation models offer a promising solution by enabling keyframe interpolation, which can enrich contextual cues and improve pose estimation performance. However, existing video generation models often lack 3D consistency, producing temporally plausible but spatially inconsistent frames that degrade downstream pose estimation. In this paper, we propose a framework **ExPose** that directly addresses 3D inconsistency when applying video generation to pose estimation in extreme-view settings. Specifically, we fine-tune a video generation model using Group Relative Preference Optimization (GRPO), aligning its outputs with 3D-consistent supervisory signals derived from pose estimation objectives. Our approach not only enhances the quality of temporal interpolation, but also ensures spatial coherence across views, significantly improving pose estimation accuracy. Extensive experiments demonstrate that our method outperforms state-of-the-art baselines, highlighting the potential of preference-optimized video generation as a powerful tool for pose estimation in extreme-view scenarios. Code is available at https://github.com/yh-yoon/ExPose.*

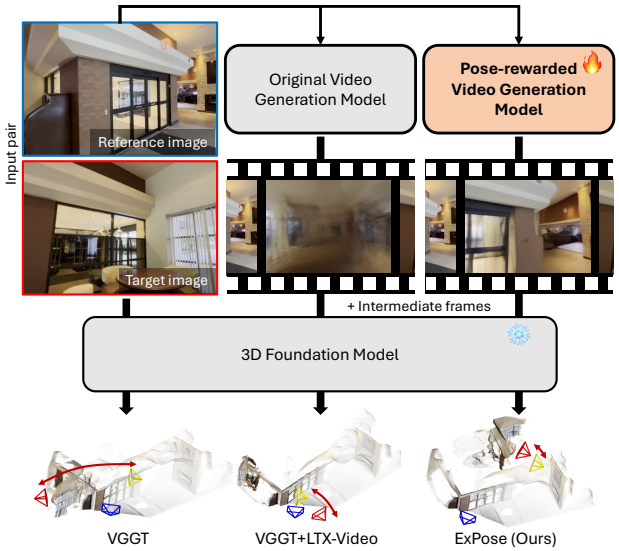

Figure 1. **Overview.** We introduce *ExPose*, a novel extreme pose estimation framework via off-the-shelf 3D foundation models with an aid of pose-rewarded reinforcement learning of video generation models. Compared to directly using the original video generation model, ExPose achieves state-of-the-art performance in extreme baseline pose estimation by synthesizing intermediate frames that more faithfully capture underlying 3D structure. Here, blue, red, and yellow denote the reference pose, the estimated target pose, and the ground-truth (GT) target pose, respectively.

1. Introduction

Pose estimation is a fundamental building block for robotics, AR/VR, autonomous driving, 3D reconstruction and even emerging tasks such as video generation. Previously, It has been tackled through iterative non-linear optimization of re-projection errors derived from visual feature correspondences, as in structure-from-motion (SfM) [32]. Such methods achieve remarkable accuracy when the input imagery provides sufficient geometric overlap, allowing robust triangulation and global consistency. However, in scenarios with sparse or weakly overlapping views, these conventional optimization-based methods often fail catastrophically, unable to recover meaningful scene geometry.

Recent progress in 3D foundation models [18, 39, 40, 42] has demonstrated that meaningful pose estimation is possible even under minimal viewpoint overlap. These models suggest that data-driven priors can complement or even replace purely geometric constraints. However, current 3D foundation models remain limited by their reliance on point-wise supervisory signals, which fail to capture the *contextual knowledge* inherent in real-world 3D environments. In contrast, humans can readily infer plausible scene layouts from only a few disjoint views, leveraging prior experience of spatial organization and object relationships. Motivated by this observation, recent research [3, 53, 55] has begun to leverage generative models trained on large-scale visual data, thereby embedding higher-level contextual understanding into the reconstruction process.

*Work done during an internship at NAVER LABS.

Trained on large-scale video datasets, video generation models [17, 29, 30, 38, 49] implicitly capture how objects, lighting, and viewpoints change in a physically and semantically consistent manner, enabling them to synthesize temporally coherent and visually realistic frame sequences that interpolate plausible transitions even between distant or sparsely related views. These capabilities make them especially well-suited for bridging the gaps between sparse input images and providing richer observations that support 3D reasoning. However, since their training objectives primarily emphasize visual realism and temporal smoothness, they lack explicit awareness of underlying 3D geometry, often producing interpolated frames that are temporally consistent yet geometrically implausible or spatially distorted. As a result, geometry-sensitive tasks such as pose estimation depend heavily on the randomness of generative videos, requiring repeated sampling and careful selection of outputs that happen to exhibit geometry-consistent behavior.

In this paper, we introduce *ExPose*, a framework designed to enhance pose estimation for 3D foundation models under extreme-view input conditions. As shown in Fig. 1, we leverage video generation models to synthesize auxiliary intermediate frames between widely separated input views, thereby enriching the contextual knowledge. To achieve this goal, we first construct a pseudo video dataset by incorporating geometrically consistent intermediate frames between sparse input images and use it for supervised fine-tuning, enabling the model to gain an initial understanding of physically plausible spatial and temporal consistency. We then further refine the model through reinforcement learning, where the pose estimation model serves as a geometry aware reward provider. Also, we adopt Group Relative Policy Optimization (GRPO) [34] to optimize the video generation model using pose-based reward signals derived from the 3D foundation model. During this process, we introduce a regularization term to promote smooth and natural transitions, and employ point-tracking based exploration to ensure diverse trajectory generation, thereby preventing overfitting to limited trajectories of camera poses.

This training paradigm enables the video generation model to learn 3D consistent motion and structure priors without requiring explicit ground-truth supervision. Once fine-tuned, the model produces intermediate frames that faithfully preserve scene geometry even under highly sparse input conditions. These generated frames are then fed into the pose estimation model, which leverages the enriched observations to perform more accurate and stable camera pose estimation. Through this synergistic integration of video generation model, ExPose bridges the gap between visual realism and spatial consistency, significantly improving pose estimation performance under challenging extreme-view scenarios.

2. Related Works

Pose Estimation Conventional structure-from-motion (SfM) [32] and visual SLAM [27, 35] jointly optimize 3D structure and camera poses from dense pixel correspondences. To further exploit reconstructed 3D scenes, neural representations [16, 26] are used to recover the scene structure and camera poses jointly [7, 47]. However, these methods are highly sensitive to photometric errors and require sufficient view overlap. To handle sparse views robustly, 3D foundation models [14, 39, 40, 42] employ feed-forward networks trained on densely annotated datasets to directly predict scene geometry and camera poses. While such models reduce the reliance on explicit optimization, their dependence on supervised data constrains scalability and generalization in extreme view scenarios. To mitigate the supervision bottleneck, recent approaches [3, 53, 55] estimate camera poses using intermediate frames generated by video generation models [17, 29, 30, 38, 49] trained on diverse large-scale video datasets. However, the generated videos often exhibit geometric inconsistencies due to the lack of explicit 3D structure, which in turn degrades pose estimation accuracy. Our method fine-tunes video generation models within a reinforcement learning framework guided by 3D foundation models to enforce pose and scene consistency, yielding structurally coherent videos that better support pose estimation under extreme views.

Video Generation Models Recent advances in video generation have been improved by large-scale diffusion transformers [29, 30, 38, 49], high-capacity VAEs [6, 17, 20, 54], and improved video captioning systems [1, 11, 51], resulting in higher visual fidelity and temporal coherence. In parallel, keyframe-conditioned approaches [4, 9, 50, 52] further demonstrate that external cues, such as reference poses or images, can guide video trajectories. Beyond these architectural developments, recent work shows that generative video models exhibit Chain-of-Frame (CoF) reasoning [8, 44], analogous to Chain-of-Thought in language models [43], enabling zero-shot spatiotemporal inference. Although video models exhibit CoF reasoning, guiding them toward a user-defined scenario remains challenging. Specifically, missing intermediate 3D information makes pose-transition videos geometrically inconsistent. To generate 3D consistent videos, we adopt a keyframe-conditioned video model [9] and introduce an RL-based fine-tuning strategy that encourages pose-aligned video generation supported by 3D foundation models. The resulting videos exhibit more coherent 3D structure, allowing 3D foundation models to estimate camera poses more reliably.

Reinforcement Learning To align large language models with human preferences, reinforcement learning has become a central approach [28, 33, 36, 56]. Preference-based objectives such as direct preference optimiza-

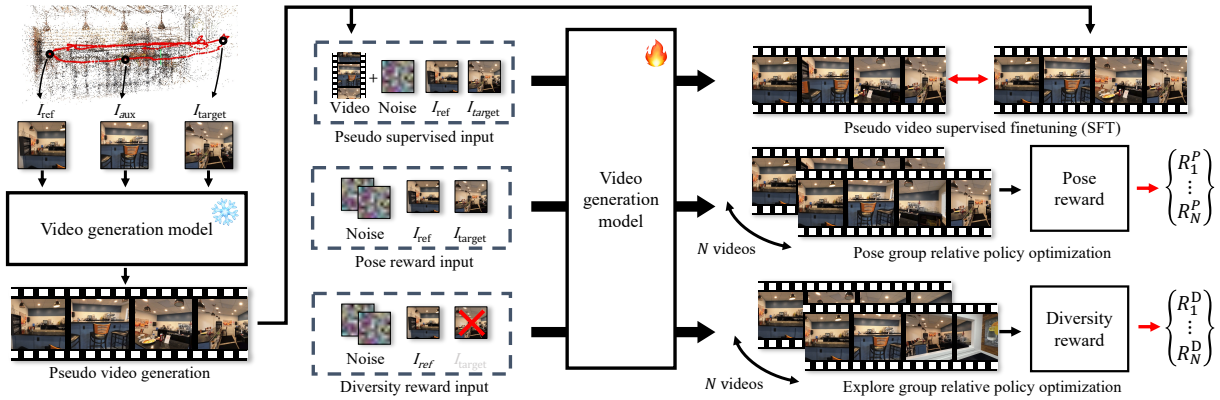

Figure 2. **Pipeline overview**. During training, an auxiliary frame I_{aux} is selected between I_{ref} and I_{target} to provide pseudo-video supervision. Then, the video generator is optimized using three components: supervised finetuning with the pseudo generated video (Sec. 3.1), an online RL strategy that employs GRPO guided by pose scores from a pretrained 3D estimation model (Sec. 3.2), and a trajectory-exploration diversity reward using only the reference frame (Sec. 3.3). During inference, the model only takes $\{I_{\text{ref}}, I_{\text{target}}\}$ as input.

tion (DPO) [31] and group relative policy optimization (GRPO) [34] further demonstrate the effectiveness of learning from pairwise or group-wise preference signals. This paradigm has expanded into generative models [2, 21, 37, 48], where video generation models incorporate reinforcement learning to enhance perceptual quality and temporal coherence [24, 25, 41, 45]. Recent work also extends GRPO training strategy to flow-matching models by reformulating deterministic flows into stochastic trajectories, enabling online policy updates in the flow space [10, 19, 23]. However, existing approaches mainly target visual or semantic preferences and lack structured geometric supervision. Our work addresses this gap by applying a GRPO-based online RL framework to a flow-matching video model, using rewards derived from 3D vision foundation models to produce videos that better preserve 3D structure and improve pose estimation under extreme view conditions.

3. Methods

In this section, we present ExPose, a framework for estimating relative camera pose under extreme-baseline conditions using only a reference–target image pair. ExPose leverages a video generator to synthesize geometry-consistent intermediate frames, enabling more reliable pose estimation where direct two-view pose inference often fails.

Our approach combines three complementary training components, and an overview of the pipeline is provided in Fig. 2. First, we incorporate supervised finetuning (Sec. 3.1) using pseudo videos generated with an auxiliary frame. This stabilizes training and helps the model retain scene structure while operating from two images. Second, we introduce pose-guided online reinforcement learning (Sec. 3.2), where GRPO updates are driven by rewards computed from pretrained 3D foundation models, encouraging geometrically consistent synthesis without requiring 3D supervision. Lastly, we add a trajectory-diversification objective based on point tracking (Sec. 3.3), encouraging the generator to explore plausible camera paths and produce diverse motion patterns.

3.1. Supervised Finetuing with Pseudo Videos

Given a reference–target input image pair $\{I_{\text{ref}}, I_{\text{target}}\}$, the pose estimator $\mathcal{F}_\theta$ outputs the target relative pose $(\hat{\mathbf{R}}, \hat{\mathbf{t}})$:

$$(\hat{\mathbf{R}}, \hat{\mathbf{t}}) \;=\; \mathcal{F}_\theta(I_{\text{ref}},\, I_{\text{target}})\,. \tag{1}$$

During training, we augment supervision by selecting an intermediate view I_{aux} between the pair and using it solely as an auxiliary frame for data construction. This yields a triplet $\{I_{\text{ref}}, I_{\text{aux}}, I_{\text{target}}\}$ that conditions pseudo-video generation and provides geometry- and video-consistency signals. At inference, only $\{I_{\text{ref}}, I_{\text{target}}\}$ are used.

To strengthen supervision under extreme baselines, we construct training triplets $\{I_{\text{ref}}, I_{\text{aux}}, I_{\text{target}}\}$ from the DL3DV dataset [22], where I_{aux} has meaningful overlap with both endpoints. We select I_{aux} among candidate frames between I_{ref} and I_{target} by subsampling a small set and retaining the one that consistently yields better downstream pose estimation performance in validation. The choice is guided by observed gains in *pose estimation accuracy* from a 3D vision foundation model. This auxiliary view anchors the in-between content and mitigates implausible scene synthesis when conditioning a video generator on sparse inputs.

Given the triplet, a multi-frame conditioned video generator produces a N-frame pseudo video

$$\mathcal{V}(I_{\text{ref}}, I_{\text{aux}}, I_{\text{target}}) \;=\; \big\{V^{(n)}(I_{\text{ref}}, I_{\text{aux}}, I_{\text{target}})\big\}_{n=1}^{N}, \tag{2}$$

where $V^{(n)}(\cdot)$ denotes the n-th frame in the pseudo video generated by pretrained video generation model. In parallel, our video generation model G_ϕ predicts a video sequence

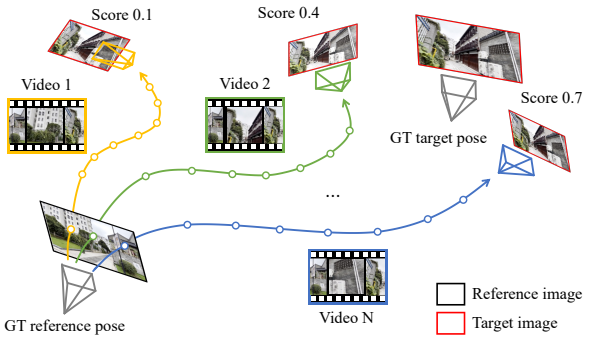

Figure 3. **Pose Rewards for Flow-GRPO [23]**. Given a reference–target image pair, multiple candidate videos are generated and evaluated with a pose estimator. Samples in each group with predicted poses closer to the ground truth receive higher rewards.

from only the reference and target pair:

$$\{\hat{V}^{(n)}(I_{\text{ref}}, I_{\text{target}})\}_{n=1}^{N} = G_\phi(I_{\text{ref}}, I_{\text{target}}) \tag{3}$$

Compared to conditioning on only two frames, incorporating I_{aux} suppresses discontinuities and unrealistic world changes that would otherwise harm pose estimation.

We adopt a single video reconstruction loss that compares with the pseudo video on a per-frame basis:

$$\mathcal{L}_{\text{SFT}} = \frac{1}{N}\sum_{n=1}^{N} \left\| \hat{V}^{(n)}(I_{\text{ref}}, I_{\text{target}}) - V^{(n)}(I_{\text{ref}}, I_{\text{aux}}, I_{\text{target}}) \right\|_1 \tag{4}$$

Here, $\hat{V}^{(n)}$ denotes the output of the video model aligned to the n-th pseudo video frame $V^{(n)}$.

3.2. Pose-guided Online Reinforcement Learning

We introduce our online RL pipeline that adapts the video generator G_ϕ to synthesize auxiliary in-between content tailored for extreme pose estimation. We use VGGT [39] to calculate the reward score of candidate pseudo videos with a pose error and optimize G_ϕ via group relative policy optimization (GRPO). We first describe how multiple candidates are produced by stochastic sampling from the Video Rectified-Flow (RF) model, then detail how VGGT-based rewards drive Flow-GRPO [23] pipeline, and finally discuss a pose interpolation constraint used for trajectory regularization.

Stochastic Sampling for Video RF model We use LTX-Video [9], a Rectified-Flow (RF) based video generation model. RF sampling is inherently deterministic—solving an ODE yields a single outcome for a given condition—so stochastic sampling is required to enable exploration in online RL and to produce multiple candidates for GRPO. Therefore, we convert the original ODE into an SDE that preserves the same marginal distribution at every time t. The deterministic RF update is

$$dx_t = v_\phi(x_t, t)\, dt \tag{5}$$

where $v_\phi(x_t, t)$ denotes the network-parameterized velocity field. We replace with the following SDE-style update to sample stochastic candidates:

$$\mathcal{D}_\phi(x_t, t) := v_\phi(x_t, t) + \tfrac{\sigma_t^2}{2t}\big(x_t + (1-t)\, v_\phi(x_t, t)\big), \tag{6a}$$

$$x_{t+\Delta t} = x_t + \mathcal{D}_\phi(x_t, t)\, \Delta t + \sigma_t \sqrt{\Delta t}\, \varepsilon \tag{6b}$$

where $\mathcal{D}_\phi(x_t, t)$ is the drift term from (6a), and σ_t controls the exploration strength; setting $\sigma_t = 0$ recovers the original deterministic RF sampling. With this ODE→SDE conversion, we can generate *diverse* video candidates under the same reference-target conditioning, which are then used for GRPO in the next phase.

Flow-GRPO with Pose Estimation Rewards We now guide the video generator G_ϕ toward in-between content that benefits relative pose accuracy. As shown in Fig. 3, we introduce a pose-derived score as feedback and update the policy with group relative policy optimization. Rather than relying only on visual plausibility, this pipeline aligns supervision with the inference goal since the generator is encouraged to produce candidates that are helpful for the downstream pose estimator $\mathcal{F}_\theta$.

For each conditioning on the same reference and target frames, we generate a group of K video candidates $\{\mathcal{V}_i\}_{i=1}^{K}$ by stochastic sampling from the Video RF model. For supervision, we obtain the ground-truth relative pose between the reference and target frames, represented by the rotation $\mathbf{R}^\star$ and the unit translation direction $\mathbf{u}^\star$. Let $(\hat{\mathbf{R}}_i, \hat{\mathbf{t}}_i)$ be the pose predicted from $\{I_{\text{ref}}, I_{\text{target}}\}$ under candidate i, and define the normalized translation direction $\tilde{\mathbf{t}}_i = \hat{\mathbf{t}}_i / \|\hat{\mathbf{t}}_i\|_2$. We compute a compact pose estimation reward that is invariant to scale and balances rotation and translation direction:

$$r_{\text{pose},i} = -\lambda_{\text{rot}}\, d_{\text{SO(3)}}(\hat{\mathbf{R}}_i, \mathbf{R}^\star) - \arccos\big(\tilde{\mathbf{t}}_i^\top \mathbf{u}^\star\big)$$

where $d_{\text{SO(3)}}$ is the geodesic rotation distance and the angular term measures unit translation direction mismatch. λ_{rot} is the scale factor of the rotation term. The reward can be averaged over multiple evaluation passes if stochastic pose estimates are used, which stabilizes training. Within each group we normalize rewards with $\bar{r} = \frac{1}{K}\sum_{k=1}^{K} r_k$ and set $s_i = r_i - \bar{r}$. Policy updates prefer higher scored samples inside the same group using a relative objective

$$\mathcal{L}_{\text{GRPO}} = -\sum_{\text{groups}} \sum_{(i \succ j)} \log \sigma\big(\beta\,(s_i - s_j)\big) + \mathcal{L}_{\text{KL}} \tag{7}$$

where σ is the logistic function and β is a temperature. An optional regularizer keeps the updated policy close to the pretrained generator to mitigate mode shrinkage:

$$\mathcal{L}_{\text{KL}} = \lambda_{\text{KL}}\, \text{KL}\big(\pi_\phi(\mathcal{V} \mid c) \,\|\, \pi_{\phi_0}(\mathcal{V} \mid c)\big),$$

where $c = (I_{\text{ref}}, I_{\text{target}})$ denotes the conditioning pair and ϕ_0 is the pretrained initialization. Here, $\pi_\phi(\mathcal{V} \mid c)$ represents

the distribution over generated video trajectories $\mathcal{V}$ under the current generator parameters. We apply the preference loss and the regularizer to update only the video generator G_ϕ while the pose estimator remains fixed. This procedure repeats online for each batch. Given a fixed pair of reference and target images we sample candidates, score them with the pose estimation reward, compute relative preferences inside the group, and update the generator.

Pose Interpolation Constraint We complement the pose estimation reward with a simple constraint that encourages camera poses inferred from each video candidate to follow a one-take video with a continuous camera trajectories. From per-frame pose estimates we extract camera centers $\{\mathbf{c}_t\}_{t=1}^{T}$ in $\mathbb{R}^3$. Let $\mathbf{c}_1$ and $\mathbf{c}_T$ be the centers of the reference and target frames and let $\mathbf{c}_m$ denote the center of a middle frame. We measure how close the middle center is to being equidistant from the endpoints, and use the deviation as a penalty so that jump cuts and fragmented trajectories receive low scores. Define $d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_2$ and $D = d(\mathbf{c}_T, \mathbf{c}_1)$. Then, the pose interpolation constraint reward is

$$r_{\text{pic}} \;=\; -\lambda_{\text{pic}} \cdot \frac{\big|\, d(\mathbf{c}_m, \mathbf{c}_1) - d(\mathbf{c}_T, \mathbf{c}_m) \,\big|}{D + \varepsilon},$$

with a small $\varepsilon > 0$ to avoid division by zero. This scale-normalized reward favors videos whose inferred camera centers vary smoothly between the endpoints and penalizes multi-take or discontinuous motion.

3.3. Exploration with Diversity Reward

For RL to be effective, the policy must explore multiple candidates per input and discover directions with higher rewards. Pure noise injection often collapses to similar trajectories. We therefore introduce an additional diversity reward that quantifies the spread of camera paths, so the model actively explores different motions.

As shown in Fig. 4, we condition only on the reference frame and encourage the generator to open diverse early trajectories. By removing the target frame from the conditioning set, the generator is free to expand into multiple plausible motions, enabling broader exploration of the motion space. For each generated video, we track a grid of points in the early segment with CoTracker [13] and define the per-video relative displacement from the first frame coordinate $\mathbf{p}_1^{(b)}(n)$ to the last frame in the early segment $\mathbf{p}_L^{(b)}(n)$ as

$$\mathbf{r}^{(b)}(n) = \mathbf{p}_L^{(b)}(n) - \mathbf{p}_1^{(b)}(n).$$

Let $v^{(b)}(n) \in \{0, 1\}$ denote visibility at the last frame. For a pair of videos (i, j), define the common visible set $S_{ij} = \{\, n \mid v^{(i)}(n) = 1 \wedge v^{(j)}(n) = 1 \,\}$. The pairwise diversity is the average L_2 distance between relative displacements over S_{ij}:

$$D_{ij} = \frac{1}{\max(1, |S_{ij}|)} \sum_{n \in S_{ij}} \|\mathbf{r}^{(i)}(n) - \mathbf{r}^{(j)}(n)\|_2,$$

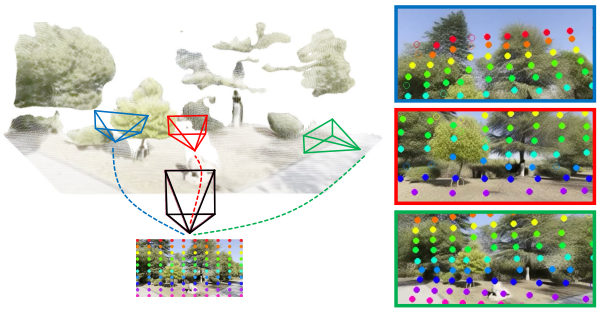

Figure 4. **Diversity reward.** Given the reference frame, we track points across generated trajectories and reward samples that exhibit larger motion diversity. This encourages our model to explore a wider range of camera paths in the early stage.

and we set $D_{ij} = 0$ if $|S_{ij}| = 0$. The per-video diversity reward is the off-diagonal row mean

$$r_{\text{div}}(i) \;=\; \lambda_{\text{div}} \cdot \frac{1}{B-1} \sum_{j \neq i} D_{ij}$$

where B denotes the number of generated video samples for a same input. This reward directly measures the spread of early trajectories. In practice, we subsample a short early window to reduce cost and use the last-frame visibility mask to aggregate points robustly.

Training Loss We train G_ϕ with a composite objective that combines supervised signals and geometry-aware preferences:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{GRPO}} \;+\; \lambda_{\text{SFT}}\, \mathcal{L}_{\text{SFT}}, \tag{8}$$

where $\mathcal{L}_{\text{SFT}}$ supervises rotation and translation with pseudo-video–induced cues (Eq. (4)), and $\mathcal{L}_{\text{GRPO}}$ encodes relative preferences guided by geometric rewards (Eq. (7)). For details about the training protocol and hyperparameter settings, please refer to the supplemental document.

4. Experiments

4.1. Experiment Settings

Dataset We evaluate extreme-baseline relative pose estimation on four datasets. Cambridge Landmarks [15] contains outdoor, scene-scale videos of streets and building facades in Cambridge. The motions are rotation-dominant, yielding low-overlap reference-target pairs that are particularly challenging for pose estimation. ScanNet [5] provides indoor, scene-scale videos of diverse environments with cluttered geometry and repeated textures, making extreme-viewpoint pairs difficult to align. NAVI [12] is an object-centric collection of videos and multiview images captured under various devices and conditions. It offers consistent object appearance across viewpoints that is beneficial for pose estimation. DL3DV [22] is a large-scale collection of scene-scale, center-facing videos captured across many points of interest, featuring complex background and camera motion

Table 1. **Quantitative comparison on the DL3DV, NAVI, and ScanNet datasets using VGGT.** We evaluate the camera pose estimates from VGGT based on images generated by each video generation model. The intermediate images produced by generative models lead to more accurate pose estimation in extreme-view scenarios, highlighting the value of high-quality generative priors. Our method performs well across most metrics on all datasets and achieves state-of-the-art results on every metric of the DL3DV dataset.

Dataset	Method	MRE ↓	MTE ↓	5°		15°		30°		
				R_{acc} ↑	T_{acc} ↑	R_{acc} ↑	T_{acc} ↑	R_{acc} ↑	T_{acc} ↑	AUC ↑
DL3DV [22]	VGGT [39]	54.28	29.08	50.00	27.33	60.00	46.33	61.67	62.33	39.79
	DynamiCrafter [46]	46.71	26.74	48.00	31.00	63.00	49.00	67.00	64.67	41.59
	Aether [55]	43.05	25.44	47.33	32.33	65.00	51.00	69.67	65.00	42.63
	LTX-Video [9]	44.13	24.32	54.33	37.00	66.33	53.00	68.67	69.00	46.88
	InterPose [3]	45.22	23.51	56.33	37.00	66.67	54.67	68.33	69.00	48.13
	ExPose (Ours)	**33.78**	**20.50**	**60.67**	**42.67**	**73.67**	**59.67**	**75.67**	**74.00**	**53.64**
NAVI [12]	VGGT [39]	27.29	14.97	43.67	52.67	76.00	79.67	79.67	83.00	63.27
	DynamiCrafter [46]	17.83	9.97	42.67	62.00	84.00	84.67	88.00	91.00	69.31
	Aether [55]	13.60	9.36	46.67	47.67	89.00	86.67	**92.67**	**93.67**	70.44
	LTX-Video [9]	18.85	10.19	47.00	62.67	81.00	86.33	88.33	89.67	69.03
	InterPose [3]	17.30	9.48	50.67	67.33	85.00	87.00	88.33	89.33	72.14
	ExPose (Ours)	**13.10**	**7.75**	**53.00**	**73.00**	**89.33**	**91.33**	91.67	93.00	**75.37**
ScanNet [5]	VGGT [39]	45.36	30.81	37.33	19.00	54.33	39.00	60.33	58.00	30.83
	DynamiCrafter [46]	42.92	31.44	32.67	13.33	54.00	27.67	51.67	46.33	19.89
	Aether [55]	40.31	30.10	31.67	16.67	56.33	38.33	64.67	60.67	30.93
	LTX-Video [9]	43.21	30.20	33.67	**20.00**	53.33	36.67	61.00	56.67	30.42
	InterPose [3]	37.72	28.87	40.33	17.67	**62.67**	40.67	67.00	61.67	34.11
	ExPose (Ours)	**36.94**	**27.04**	**41.67**	17.67	61.00	**42.00**	**67.33**	**66.00**	**34.73**

Table 2. **Quantitative comparison on the Cambridge Landmarks dataset using VGGT.** Our method achieves state-of-the-art performance across all rotation error metrics.

Cambridge Landmarks [15]					
Method	MRE ↓	R_{acc} ↑			AUC ↑
		5°	15°	30°	30°
VGGT [39]	17.97	59.67	78.33	82.33	71.09
DynamiCrafter [46]	20.17	56.00	75.67	80.67	68.43
Aether [55]	19.28	60.00	79.00	84.33	68.00
LTX-Video [9]	15.40	62.67	81.67	87.00	74.60
InterPose [3]	15.36	66.33	80.00	86.67	74.51
ExPose (Ours)	**11.48**	**72.00**	**86.00**	**90.00**	**79.44**

that enable rigorous extreme-baseline evaluation. For each dataset, we form evaluation pairs by selecting reference and target images with small visual overlap and significant viewpoint change. The selected pairs will be released for reproducibility.

Pose Estimation Models We assess downstream effectiveness using two pose estimators. VGGT [39] is a transformer-based 3D foundation model for multi-view geometry. MapAnything [14] is a geometry-aware estimator that predicts relative rotation and translation direction.

Video Generation Models We compare our generator against four video generation models. DynamiCrafter [46] is a diffusion-based video model for dynamic content. Aether [55] is a high-fidelity video model. LTX-Video [9], which we adopt as our backbone, is a rectified-flow video generator. InterPose [3] proposes a test-time scaling strategy that improves pose estimation by generating multiple video candidates and selecting samples that yield consistent relative poses. For a fair comparison, we reimplement InterPose [3] with LTX-Video and follow the original setting by producing four scaled samples at test time.

Evaluation Metrics We report mean rotation error (MRE) and mean translation direction error (MTE), along with percentage rotation (R_{acc}) and translation (T_{acc}) errors within the 5°/15°/30° thresholds. We also include the area under the accuracy curve (AUC) up to 30°. All metrics are computed on the same set of image pairs for each dataset.

4.2. Quantitative Analysis

We evaluate how different video generators influence extreme-baseline pose estimation by fixing the pose estimator (VGGT [39] or MapAnything [14]) and comparing the intermediate frames produced under identical reference–target conditioning and computational budgets.

With VGGT as the estimator (Tabs. 1 and 2), our method yields the best performance across all benchmarks. On DL3DV, NAVI, and ScanNet, where viewpoint shifts are large and overlap is low, our generated frames consistently improve both rotation accuracy and translation-direction metrics over all baselines and over VGGT without generated frames. Even on the rotation-dominant Cambridge Landmarks dataset, our method remains superior, indicating that the proposed training stages enhance geometric consistency between the input views.

Table 3. **Quantitative comparison on the DL3DV, NAVI, and ScanNet using MapAnything.** We evaluate camera pose estimations from MapAnything using images generated by different video generation models, where intermediate frames help improve accuracy in extreme-view settings. Our method further achieves state-of-the-art performance across all metrics on the NAVI and ScanNet datasets.

Dataset	Method	MRE ↓	MTE ↓	5°		15°		30°		
				R_{acc} ↑	T_{acc} ↑	R_{acc} ↑	T_{acc} ↑	R_{acc} ↑	T_{acc} ↑	AUC ↑
DL3DV [22]	MapAnything [14]	35.59	25.62	55.00	36.33	**71.00**	50.33	73.67	65.67	43.60
	DynamiCrafter [46]	36.88	25.54	53.00	35.67	67.33	51.33	72.00	65.33	41.66
	Aether [55]	36.19	25.23	53.33	35.67	68.33	54.00	72.33	67.00	44.04
	LTX-Video [9]	35.30	24.50	56.33	39.00	68.67	52.00	73.33	65.67	45.42
	InterPose [3]	35.25	24.85	**58.67**	40.33	**71.00**	**55.67**	74.67	65.33	46.07
	ExPose (Ours)	**34.83**	**23.68**	**58.67**	**41.00**	70.33	**55.67**	**75.00**	**68.33**	**46.81**
NAVI [12]	MapAnything [14]	27.16	12.85	**27.00**	46.67	69.33	78.67	81.67	86.00	57.61
	DynamiCrafter [46]	26.37	12.96	22.00	42.33	67.00	78.33	81.00	87.67	55.67
	Aether [55]	24.36	12.73	23.33	43.00	71.00	79.33	82.67	88.33	57.49
	LTX-Video [9]	26.51	12.65	25.33	44.33	66.33	79.00	80.00	86.67	56.02
	InterPose [3]	24.37	12.09	**27.00**	45.67	71.33	81.00	83.67	87.67	59.00
	ExPose (Ours)	**21.99**	**11.05**	**27.00**	**47.33**	**74.67**	**82.33**	**86.00**	**89.33**	**60.87**
ScanNet [5]	MapAnything [14]	58.55	37.00	**29.33**	10.33	48.00	30.00	52.67	47.67	21.29
	DynamiCrafter [46]	59.49	37.02	23.33	6.33	46.00	27.67	51.67	46.33	19.89
	Aether [55]	57.35	36.78	20.33	6.33	47.00	30.33	52.67	45.33	20.03
	LTX-Video [9]	54.71	35.29	26.67	10.00	48.00	32.00	56.00	49.00	22.79
	InterPose [3]	57.37	34.42	28.67	12.00	**51.00**	33.33	55.00	**51.00**	23.33
	ExPose (Ours)	**52.31**	**33.28**	**29.33**	**13.67**	**51.00**	**35.33**	**57.00**	**51.00**	**25.21**

Table 4. **Quantitative comparison on the Cambridge Landmarks dataset using MapAnything.** Our method improves downstream pose estimator performance across other video generation models by producing more geometrically consistent videos.

Cambridge Landmarks [15]					
Method	MRE ↓	R_{acc} ↑			AUC ↑
		5°	15°	30°	30°
MapAnything [14]	21.46	62.00	81.00	83.67	73.32
DynamiCrafter [46]	22.48	58.00	78.33	81.33	70.14
Aether [55]	20.32	60.00	79.00	84.33	72.20
LTX-Video [9]	18.96	**67.33**	82.33	85.00	75.08
InterPose [3]	19.51	61.67	81.67	84.67	74.51
ExPose (Ours)	**16.08**	63.00	**84.67**	**87.67**	**76.52**

Using MapAnything (Tabs. 3 and 4) shows the same trend. Across all datasets, our method yields more reliable supervision than existing video models, and the improvements persist across all relative pose metrics. The agreement between the two distinct pose estimators indicates that the gains stem from the geometric quality of our generated frames rather than properties of any pose estimator.

4.3. Qualitative Comparison and Analysis

We present qualitative comparisons on the Cambridge Landmarks, ScanNet, DL3DV-10K, and NAVI datasets, evaluating the original VGGT without video supervision, VGGT enhanced with LTX-Video, and our method (Fig. 5). When the reference and target views share minimal geometric overlap, as common in both indoor and outdoor scenes, existing models struggle to reliably regress the target camera pose. In particular, VGGT often produces geometrically inconsistent predictions under large viewpoint changes, and the LTX-Video–augmented variant shows partial improvement but still fails to maintain accurate alignment. In contrast, our method consistently estimates the most accurate and geometrically coherent poses, even in scenarios with extremely sparse view overlap. These qualitative results demonstrate that our approach provides strong robustness and generalization across diverse environments, outperforming all baselines in challenging viewpoint configurations.

4.4. Ablation Studies

Effect of Model Components Table 5 summarizes the quantitative ablation results for the four key components of our framework: supervised finetuning (SFT), group relative policy optimization (GRPO), pose interpolation constraint (PIC), and diversity reward (Div). Adding SFT to the video-only baseline significantly reduces MRE and improves accuracy across all thresholds, showing that basic supervised learning on generated videos is essential. Incorporating GRPO further boosts both rotational and translational accuracy, particularly at the 15° threshold, by directly optimizing pose-based rewards. Introducing PIC enhances trajectory consistency and leads to additional reductions in MTE. Finally, our full model with the diversity reward achieves the best performance on all metrics, demonstrating that encouraging diverse viewpoint generation further strengthens overall pose estimation accuracy.

Table 5. **Quantitative ablation study.** We evaluate four key components of our method: supervised finetuning (SFT), group relative policy optimization (GRPO), pose interpolation constraint (PIC), and diversity reward (Div). The evaluation uses the DL3DV dataset, with LTX-Video for video generation and VGGT for pose estimation.

Variant	MRE ↓	MTE ↓	5°		15°		30°		
			R_{acc} ↑	T_{acc} ↑	R_{acc} ↑	T_{acc} ↑	R_{acc} ↑	T_{acc} ↑	AUC ↑
Video only	44.13	24.32	54.33	37.00	66.33	53.00	68.67	69.00	46.88
SFT	35.48	21.67	59.00	41.00	72.67	59.67	75.33	71.67	51.84
SFT+GRPO	34.41	21.35	59.67	43.33	**73.67**	60.00	75.00	72.67	52.84
SFT+GRPO+PIC	35.51	20.72	60.33	**44.00**	73.00	**60.67**	**75.67**	73.67	53.31
SFT+GRPO+PIC+Div (Ours)	**33.78**	**20.50**	**60.67**	42.67	**73.67**	59.67	**75.67**	**74.00**	**53.64**

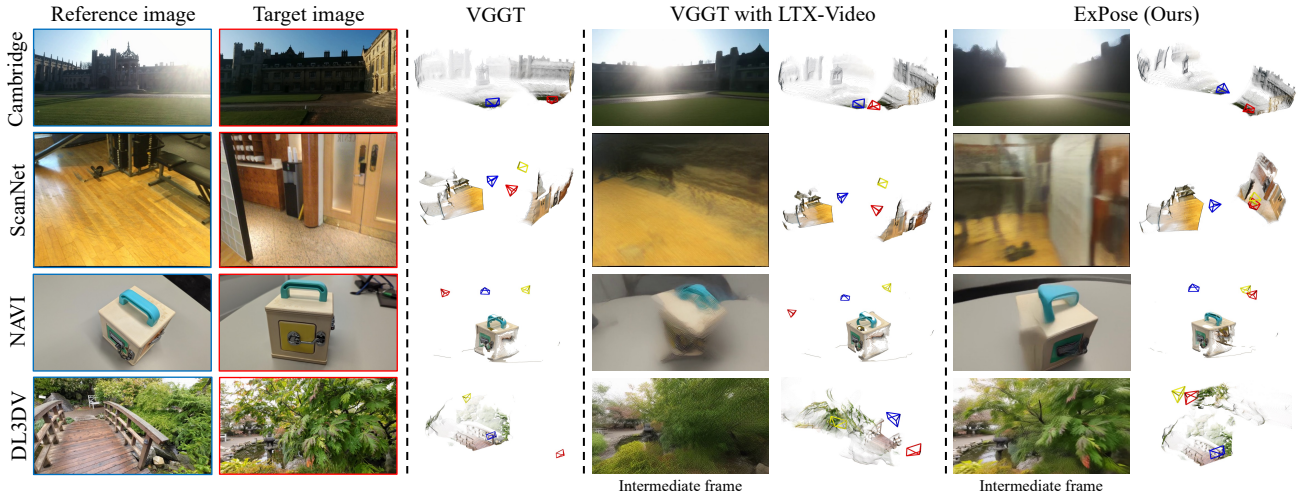

Figure 5. **Qualitative comparisons.** We show the input image pair in the first two columns: blue denotes reference image and pose, and red denotes target image and its estimated pose. In the third column we show the VGGT predicted point maps and camera poses, using the input image pair. Here, yellow denotes ground-truth (GT) target pose. In the 4th and 5th columns are shown LTX-Video generated intermediate frames and VGGT results with the additional image frames. The final two columns show corresponding results from ExPose. For the Cambridge dataset, GT poses are not visualized because only rotation ground truth is available. Note that for visualization, the scale of the GT pose is normalized by matching its distance to the reference frame to that of the predicted target frame.

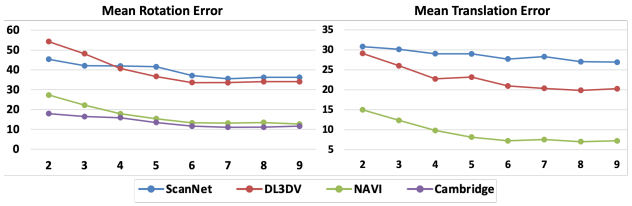

Figure 6. **Ablation study of the number of intermediate frame.** Mean pose errors (MRE and MTE) decrease with an increasing number of intermediate frames and converge beyond seven frames.

The Number of Intermediate Frames. We demonstrate the effect of the number of intermediate frames on pose estimation in Fig. 6. Both the mean rotation error (MRE) and mean translation error (MTE) consistently decrease as more intermediate frames are included, indicating that additional temporal cues help stabilize pose regression under large viewpoint differences. The performance improvements saturate beyond seven frames, suggesting that our method effectively leverages temporal information without requiring densely sampled trajectories.

5. Conclusion

We introduced ExPose, a novel framework that aligns video generation with extreme-baseline pose estimation. ExPose integrates pseudo-video supervised finetuning with online reinforcement learning guided by pose-derived rewards. By sampling multiple candidate sequences per input and refining the generator with GRPO, the method achieves effective exploration, while a pose-interpolation constraint and a point-tracking–based diversity reward deliver smooth, geometry-consistent intermediate views and prevent degenerate trajectory collapse. Collectively, these components enable ExPose to effectively integrate generative video reasoning with geometric constraints. Across four datasets and two pose estimators, ExPose achieves consistent gains in both rotation and translation accuracy, demonstrating strong robustness to severe viewpoint gaps and highlighting its potential to reshape how video generation models contribute to geometry-aware 3D vision tasks. We believe our work opens a promising direction for future works addressing the integration of generative modeling and 3D geometry.

Acknowledgments

This work was supported by the InnoCORE program of the Ministry of Science and ICT(N10250156); by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00457882, AI Research Hub Project); and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25473963).

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[2] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 3

[3] Ruojin Cai, Jason Y Zhang, Philipp Henzler, Zhengqi Li, Noah Snavely, and Ricardo Martin-Brualla. Can generative video models help pose estimation? In *CVPR*, pages 16764–16773, 2025. 1, 2, 6, 7

[4] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025. 2

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 5, 6, 7

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 2

[7] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Sparse-view gaussian splatting in seconds. *arXiv preprint arXiv:2403.20309*, 2024. 2

[8] Ziyu Guo, Xinyan Chen, Renrui Zhang, Ruichuan An, Yu Qi, Dongzhi Jiang, Xiangtai Li, Manyuan Zhang, Hongsheng Li, and Pheng-Ann Heng. Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. *arXiv preprint arXiv:2510.26802*, 2025. 2

[9] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 2, 4, 6, 7

[10] Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. *arXiv preprint arXiv:2508.04324*, 2025. 3

[11] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2

[12] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, Andre Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. NAVI: Category-agnostic image collections with high-quality 3d shape and pose annotations. In *NeurIPS*, 2023. 5, 6, 7

[13] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *ECCV*, 2024. 5

[14] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 2, 6, 7

[15] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, pages 2938–2946, 2015. 5, 6, 7

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):139–1, 2023. 2

[17] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[18] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, pages 71–91. Springer, 2024. 1

[19] Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025. 3

[20] Zongjian Li, Bin Lin, Yang Ye, Liuhan Chen, Xinhua Cheng, Shenghai Yuan, and Li Yuan. Wf-vae: Enhancing video vae by wavelet-driven energy flow for latent video diffusion model. In *CVPR*, pages 17778–17788, 2025. 2

[21] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2 (5):7, 2024. 3

[22] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, pages 22160–22169, 2024. 3, 5, 6, 7

[23] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 3, 4

[24] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia,

Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025. 3

[25] Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In *CVPR*, pages 8009–8019, 2025. 3

[26] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[27] David Nistér, Oleg Naroditsky, and James Bergen. Visual odometry. In *CVPR*. Ieee, 2004. 2

[28] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744, 2022. 2

[29] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 2

[30] Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, et al. Open-sora 2.0: Training a commercial-level video generation model in $200 k. *arXiv preprint arXiv:2503.09642*, 2025. 2

[31] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 3

[32] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1, 2

[33] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2

[34] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 3

[35] Randall C Smith and Peter Cheeseman. On the representation and estimation of spatial uncertainty. *The international journal of Robotics Research*, 5(4):56–68, 1986. 2

[36] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020. 2

[37] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, pages 8228–8238, 2024. 3

[38] 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. 2

[39] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025. 1, 2, 4, 6

[40] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, pages 20697–20709, 2024. 1, 2

[41] Shengzhi Wang, Yingkang Zhong, Jiangchuan Mu, Kai Wu, Mingliang Xiong, Wen Fang, Mingqing Liu, Hao Deng, Bin He, Gang Li, et al. Align-a-video: Deterministic reward tuning of image diffusion models for consistent video editing. In *CVPR*, pages 2074–2083, 2025. 3

[42] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 1, 2

[43] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022. 2

[44] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025. 2

[45] Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. Densedpo: Fine-grained temporal preference optimization for video diffusion models. *arXiv preprint arXiv:2506.03517*, 2025. 3

[46] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *ECCV*, pages 399–417. Springer, 2024. 6, 7

[47] Jiale Xu, Shenghua Gao, and Ying Shan. Freesplatter: Pose-free gaussian splatting for sparse-view 3d reconstruction. In *ICCV*, pages 25442–25452, 2025. 2

[48] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *CVPR*, pages 8941–8951, 2024. 3

[49] 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. 2

[50] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025. 2

[51] Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888*, 2025. 2

[52] Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *NeurIPS*, 2025. 2

[53] Qitao Zhao and Shubham Tulsiani. Sparse-view pose estimation and reconstruction via analysis by generative synthesis. *NeurIPS*, 37:111899–111922, 2024. 1, 2

[54] Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan. Cv-vae: A compatible video vae for latent generative video models. *NeurIPS*, 37:12847–12871, 2024. 2

[55] Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *ICCV*, pages 8535–8546, 2025. 1, 2, 6, 7

[56] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. 2