Distilling Geometry Priors for 3D-Consistent Video Generation

Anonymous CVPR submission

Paper ID *****

Abstract

*While recent video diffusion models (VDMs) produce visually impressive results, they fundamentally struggle to maintain 3D structural consistency, often resulting in object deformation or spatial drift. We hypothesize that these failures arise because standard denoising objectives lack explicit incentives for geometric coherence. To address this, we introduce **VideoGPA** (**Video** **G**eometric **P**reference **A**lignment), a data-efficient self-supervised framework that leverages a geometry foundation model to automatically derive dense preference signals that guide VDMs via Direct Preference Optimization (DPO). This approach effectively steers the generative distribution toward inherent 3D consistency without requiring human annotations. VideoGPA significantly enhances temporal stability, physical plausibility, and motion coherence using minimal preference pairs, consistently outperforming state-of-the-art baselines in extensive experiments.*

1. Introduction

In this work, we introduce **VideoGPA** (**Video** **G**eometric **P**reference **A**lignment), a data-efficient, self-supervised framework that equips pre-trained VDMs with 3D consistency. The key to our approach is a novel 3D consistency metric derived from the principle of re-projection consistency. Specifically, given a generated video, we utilize a GFM to render a 3D consistent video as reference. The discrepancy between the input video and the rendered reference serves as a robust proxy for 3D consistency: if a video is geometrically valid, the GFM-derived 3D structure should accurately reconstruct the original input. By leveraging a geometry foundation model as a reward model backbone, we construct geometric preference pairs, distinguishing between samples with high and low structural integrity. These pairs guide the VDM via Direct Preference Optimization (DPO) [7], effectively steering the generative distribution toward the 3D-consistent manifold. Remarkably, we show that with only $\sim 2,500$ preference pairs and minimal post-training (LoRA fine-tuning [3] on $\sim$1% of model parameters), VideoGPA substantially improves geometric coherence and temporal stability, while preserving the base model's visual quality and motion realism. Across both image-to-video and text-to-video settings, VideoGPA consistently improves over prior art on multiple geometric consistency and perceptual metrics.

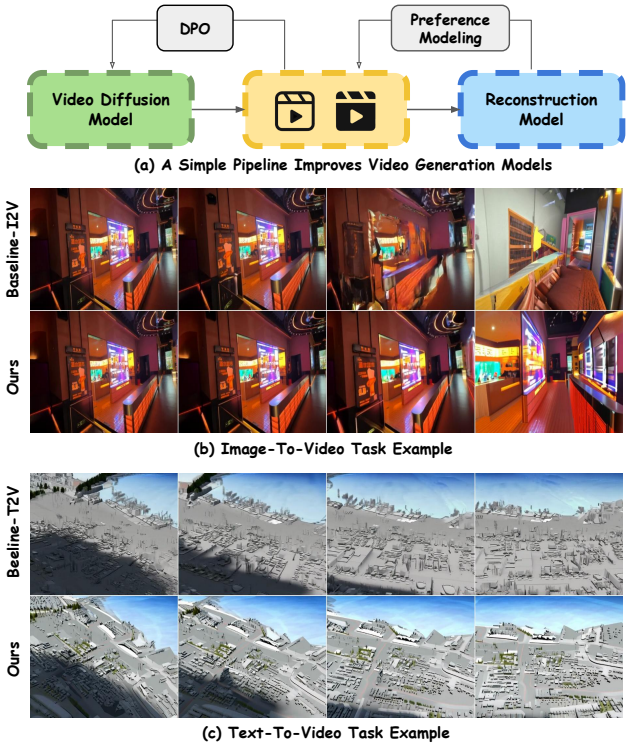

Figure 1. **Overview of VideoGPA and representative results.** (a) VideoGPA aligns a pretrained video diffusion model through proposed reconstruction-guided preference optimization. (b) Image-to-video examples comparing the base model [11] and VideoGPA, showing improved geometric stability under camera motion. (c) Text-to-video examples demonstrating improved structural coherence and reduced geometric artifacts.

2. Methodology

VideoGPA introduces a review-and-correct framework that aligns video diffusion models with 3D physical laws. As shown in Fig. 2, the process begins by using a geometry

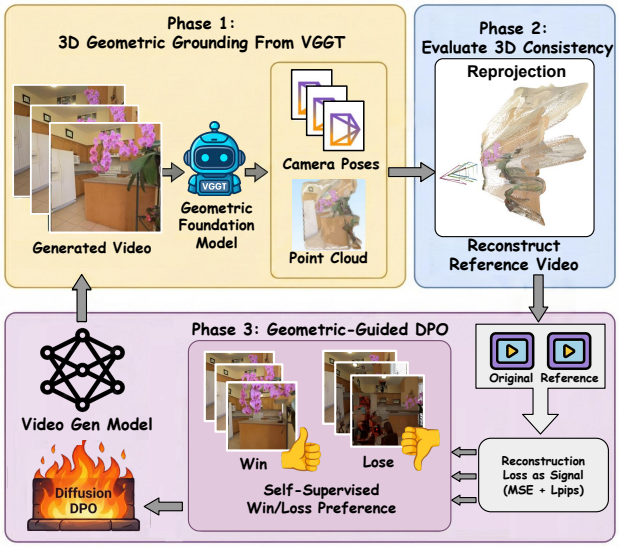

Figure 2. **Pipeline of VideoGPA.** A geometric foundation model probes generated videos to assess scene-level 3D consistency, which is used to form self-supervised preference pairs for post-training alignment via DPO.

foundation model to extract the 3D structure and camera motion from generated videos. We then calculate a 3D consistency score by measuring the error in reconstructing the original frames from this 3D structure. Finally, with preference pairs constructed using these scores, we apply DPO to teach the model to favor geometrically consistent outputs. This lightweight approach allows the model to maintain rigid structures with minimal additional training.

2.1. DPO for v-Prediction Video Diffusion

Recent advancements in video generation have been significantly driven by diffusion transformers (DiTs) employing the v-prediction parameterization [2, 5, 8]. Trained on large-scale datasets, these models demonstrate remarkable potential in synthesizing high-fidelity dynamic content with stable training convergence. Building upon this foundation, we adapt the Diffusion-DPO [9] framework to this parameterization to explicitly steer the model toward geometrically consistent manifolds.

For a general diffusion model trained with v-prediction, the target velocity v_t at timestep t is formally defined as:

$$v_t \equiv \dot{x}_t = \alpha_t \epsilon - \sigma_t x_0, \tag{1}$$

where α_t and σ_t are the noise schedule coefficients. The network v_θ is trained to minimize the mean squared error (MSE) relative to the velocity target v_t.

By substituting the velocity error into the DPO framework, we define the energy term $\mathcal{E}$ for a sample x as:

$$\mathcal{E}(\theta, x, t) = \|v_t - v_\theta(x_t, t, c)\|^2. \tag{2}$$

The log-probability ratio in the velocity space thus becomes:

$$\log \frac{\pi_\theta(x|c)}{\pi_{\text{ref}}(x|c)} \propto \mathbb{E}_{t,\epsilon}\left[\mathcal{E}(\text{ref}, x, t) - \mathcal{E}(\theta, x, t)\right] \tag{3}$$

The final DPO loss for our v-prediction video model is formulated as the negative log-likelihood of the reward margin between the winning and losing samples:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}\Big[\log\sigma\Big(\beta\Big([\mathcal{E}(\text{ref}, x^w, t) - \mathcal{E}(\theta, x^w, t)] - [\mathcal{E}(\text{ref}, x^l, t) - \mathcal{E}(\theta, x^l, t)]\Big)\Big)\Big]. \tag{4}$$

During training, we sample shared noise ϵ and timestep t for each preference pair (x^w, x^l) to ensure a consistent optimization baseline. The model v_θ (parameterized via LoRA [3]) is updated to minimize the velocity prediction error for x^w relative to x^l, effectively steering the latent video manifold toward higher geometric consistency.

2.2. Preference Modeling

DPO requires preference pairs to guide post-training alignment. In our setting, the objective favors video samples that exhibit stronger 3D geometric consistency. To this end, we derive a self-supervised geometric preference signal by analyzing generated videos with a feed-forward geometric foundation model [10]. The resulting 3D consistency score enables automatic construction of preference pairs without human annotations or explicit structural priors. We next describe how this score is computed from reconstructed camera motion and scene geometry.

For each generated video, we uniformly sample T frames to obtain an image sequence $\mathcal{I} = \{I_t\}_{t=1}^T$. Given this sequence, the geometric model Φ predicts a depth map D_t and camera pose (R_t, t_t) for each frame I_t,

$$(D_t, R_t, t_t) = \Phi_\theta(I_t), \qquad R_t \in SO(3),\ t_t \in \mathbb{R}^3, \tag{5}$$

along with camera intrinsics $K \in \mathbb{R}^{3\times3}$. Using the camera-to-world transform $E_t = [R_t|t_t] \in SE(3)$, each pixel $\tilde{\mathbf{u}} = [u, v, 1]^\top$ with depth $D_t(u, v) > 0$ can be projected to the world coordinate frame with:

$$\begin{aligned} \mathbf{x}_t^{\text{cam}}(u, v) &= D_t(u, v)\, K^{-1}\tilde{\mathbf{u}}, \\ \mathbf{X}_t(u, v) &= R_t\, \mathbf{x}_t^{\text{cam}}(u, v) + t_t, \end{aligned} \tag{6}$$

formulating a colored point cloud $\mathcal{P} = \{(\mathbf{X}_i, \mathbf{c}_i)\}_{i=1}^N$, where $\mathbf{c}_i$ is RGB color. By default, we set $T = 10$.

3D Consistency Score We measure 3D geometric consistency by how well the recovered structure explains the original frames under reprojection. Geometrically coherent videos admit a consistent 3D explanation across viewpoints, while inconsistencies lead to elevated reprojection error.

Table 1. **Quantitative evaluation** on image-to-video (I2V) and text-to-video (T2V) generation. We report 3D reconstruction error, 3D geometric consistency metrics, and human-aligned VideoReward scores. Overall, VideoGPA consistently achieves the strongest geometric consistency while maintaining or improving perceptual quality across both I2V and T2V settings.

Method	3D Reconstruction Error			3D Consistency			VideoReward (Win Rate %)			
	PSNR ↑	SSIM ↑	LPIPS ↓	MVCS ↑	3DCS ↓	Epipolar ↓	VQ	MQ	TA	OVL
Image-to-Video (I2V) *Base Model: CogVideoX-I2V-5B*										
Baseline-I2V	14.57	0.455	0.653	0.976	0.687	0.706	-	-	-	-
SFT	**15.23**	0.509	0.639	0.982	0.665	0.628	44.67	33.00	52.67	35.00
Epipolar-DPO	15.07	0.479	0.615	0.984	0.646	0.571	67.33	51.33	56.67	66.00
VideoGPA (Ours)	15.19	**0.510**	**0.608**	**0.986**	**0.638**	**0.564**	**74.00**	**56.00**	**57.67**	**76.00**
Text-to-Video (T2V) *Base Model: CogVideoX-5B*										
Baseline-T2V	17.53	0.614	0.508	0.967	0.533	0.584	-	-	-	-
SFT	17.07	0.573	0.563	0.968	0.586	0.719	14.67	23.67	39.33	15.33
Epipolar-DPO	**17.69**	0.618	0.507	0.971	0.528	0.579	45.00	53.67	49.00	48.67
VideoGPA (Ours)	17.31	**0.621**	**0.495**	**0.974**	**0.519**	**0.548**	**62.67**	**67.00**	42.67	**60.33**

Formally, each 3D point $\mathbf{X} \in \mathcal{P}$ is reprojected into frame k using inverse camera pose $E_{t,\mathrm{w2c}} = [R_t^\top \mid -R_t^\top t_t]$, yielding

$$\mathbf{x}_t^{\mathrm{cam}} = R_t^\top (\mathbf{X} - t_t), \quad (u_t, v_t) = \pi(K\,\mathbf{x}_t^{\mathrm{cam}}), \quad (7)$$

where π denotes perspective division. Using vectorized rendering with a painter's algorithm, we obtain a reprojected image $\hat{I}_t$ for each frame.

We quantify geometric consistency between reprojected image $\{\hat{I}_t\}_{t=1}^T$ and original frame $\{I_t\}_{t=1}^T$ using a standard reconstruction loss [12]:

$$E_{\mathrm{Recon}} = \frac{1}{T}\sum_{t=1}^{T}\Big(\mathrm{MSE}\big(\hat{I}_t,\, I_t\big) + \mathrm{LPIPS}\big(\hat{I}_t,\, I_t\big)\Big). \quad (8)$$

Lower reconstruction error indicates stronger cross-view geometric consistency, while higher error reflects violations of 3D coherence. We use this reprojection-based error as a dense, self-supervised signal to construct preference pairs.

3. Experiments

3.1. Experimental Results

3.1.1. Evaluation Metrics

We evaluate geometric fidelity and perceptual quality via: **(1) 3D Reconstruction Error:** PSNR, SSIM, and LPIPS between original and reprojected frames from the 3D geometry. **(2) 3D Consistency:** Multi-View Consistency Score (MVCS [1]) , our 3D Consistency Score (3DCS) , and Epipolar Sampson Error [4]. **(3) Human-Aligned Quality:** VideoReward [6] win rates covering Visual (VQ), Motion (MQ), Text Alignment (TA), and Overall (OVL) scores. Static videos are filtered to ensure evaluation on the motion manifold.

3.1.2. Quantitative Evaluation

Human Preference: A blind study with 25 participants (20 groups/person) shows VideoGPA is preferred in 53.5% of cases, doubling the performance of the next-best method, Epipolar-DPO (22.4%) (Fig. 3).

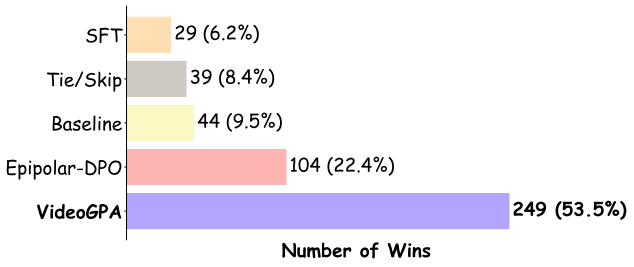

Figure 3. **Human preference study** on I2V generation. VideoGPA is most frequently preferred, indicating improved perceptual quality and 3D consistency.

4. Discussion

Scene-Level vs. Local Geometry: Unlike local metrics (e.g., Epipolar-DPO [4]) that rely on pairwise relations, VideoGPA enforces a global reprojection constraint. Local metrics are prone to *false positives* where degenerate outputs like texture collapse or frozen regions satisfy sparse constraints (Fig. 5). By requiring all frames to jointly admit a single 3D explanation, VideoGPA provides a dense, unambiguous signal that prevents spatial drift and stabilizes alignment.

Geometry as a Motion Regularizer: Although targeting static geometry, VideoGPA improves dynamic motion coherence (Fig. 6) and MQ scores (Table 1). We posit that VideoGPA acts as a *geometric regularizer*, projecting generations onto a physically plausible manifold subspace. By "fixing the stage" (stable background and camera), the model can better disentangle camera movement from object motion, allowing inherent priors to focus on coherent object dynamics.

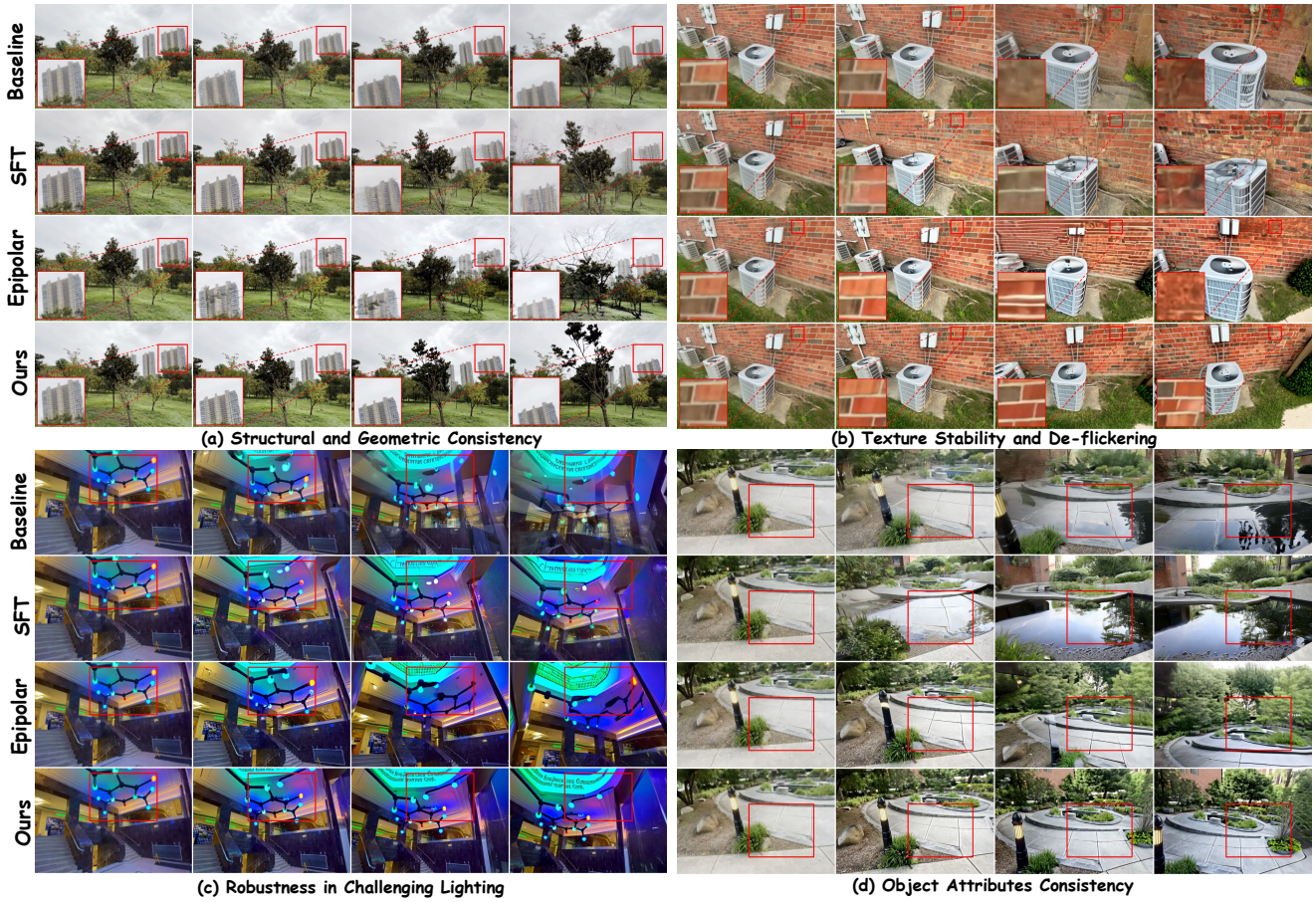

Figure 4. **Qualitative comparison** on I2V generation. We compare VideoGPA with the base model, SFT, and Epipolar-DPO. Highlighted regions illustrate improvements in (a) structural and geometric consistency, (b) texture stability and de-flickering, (c) robustness under challenging lighting, and (d) object attribute consistency.

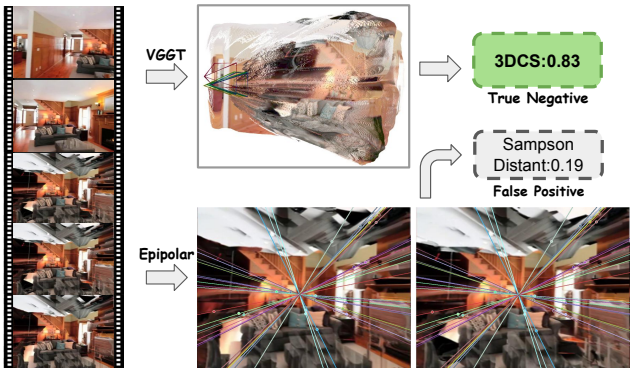

Figure 5. **Scene-level *v.s.* local geometry.** Comparison between local geometric metrics and the proposed scene-level 3D consistency score on a corrupted video. Local, pairwise constraints yield a false positive, while the scene-level metric correctly identifies geometric inconsistency.

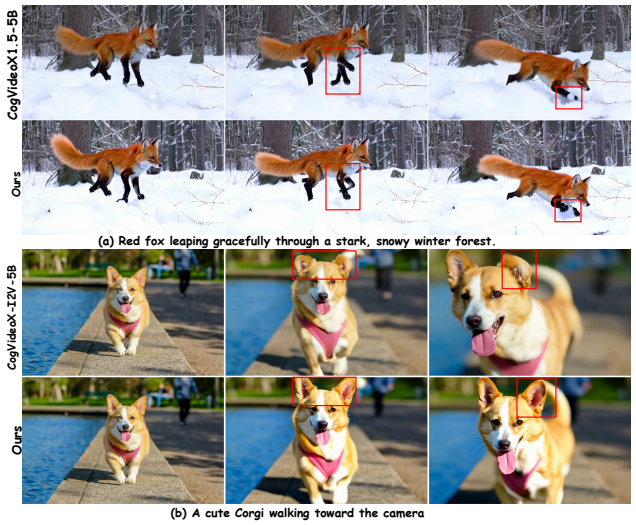

Figure 6. **Improved object motion coherence** in both T2V (top) and I2V (bottom) generation. VideoGPA better preserves object structure and motion continuity across frames.

References

[1] Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. Geovideo: Introducing geometric regular-

ization into video generation model. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[2] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[4] Orest Kupyn, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models. *arXiv preprint arXiv:2510.21615*, 2025.

[5] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[6] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[7] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[8] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[9] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[10] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[11] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[12] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.