Learning to Correct Geometry in Generated Videos

Weijie Lyu[1] **Tiancheng Shen**[1] **Jason Li**[2] **Yujing Wang**[3]
Xiaoyuan Wang[4] **Haobo Yuan**[1] **Yizhou Zhao**[4] **Ming-Hsuan Yang**[1]

[1]University of California, Merced [2]Nanyang Technological University
[3]Peking University [4]Carnegie Mellon University

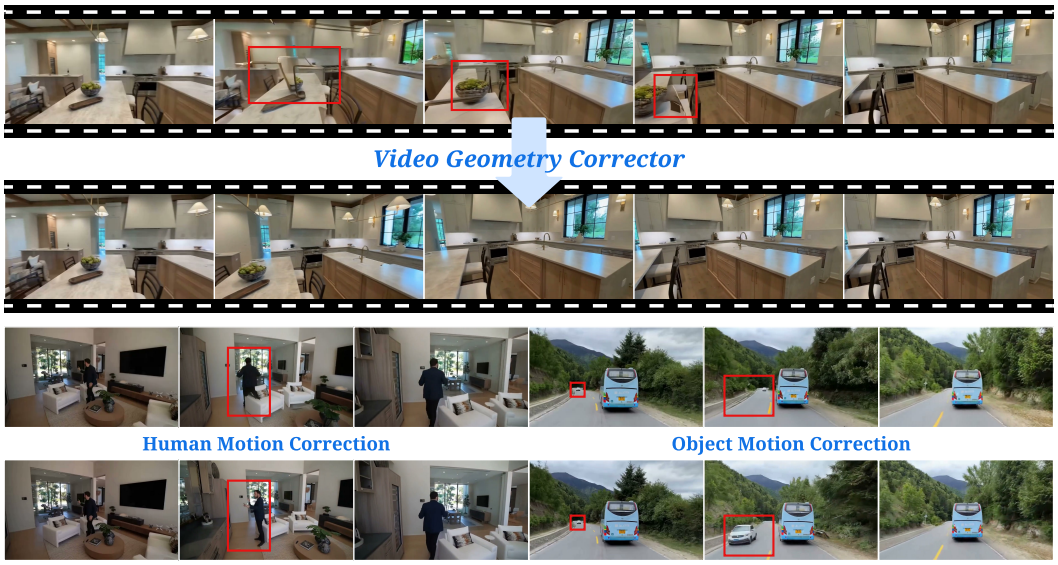


Figure 1: **Video Geometry Corrector** (*VGC*) corrects geometry distortions in generated videos. Under large camera motion, modern video generators often produce warped structures, inconsistent parallax, and drifting scene layouts (red boxes). *VGC* refines these outputs into geometry-consistent videos. When scaled to a larger backbone, it further extends to general video refinement, including dynamic scenes with distorted human or object motion. Project page: https://weijielyu.github.io/VGC.

Abstract

Generated videos often lose geometric consistency under camera motion: structures warp under viewpoint change, parallax becomes inconsistent, and smooth camera movement may collapse into shot-like transitions. We argue that this failure is largely a data problem. In web-scale video corpora, most visible motion comes from dynamic objects captured by static or weakly moving cameras, leaving only sparse supervision for viewpoint-induced appearance changes. To address this, we propose Video Geometry Corrector (*VGC*), a post-hoc video-to-video model trained on synthetic paired supervision. Starting from real videos with diverse camera motion and pose annotations, we construct *VG-Pairs* with a synthetic paired-supervision data generation pipeline. Each video clip is treated as a geometry-consistent target and paired with a distorted counterpart synthesized by a pretrained image-to-video inpainting model using trajectory-adaptive keyframes. *VGC* learns to map the distorted input to the geometry-consistent target without modifying the source generator, and uses optical flow as an auxiliary cue to help disentangle camera motion from object motion. Experiments on pose-conditioned and prompt-driven benchmarks show that *VGC* improves geometric consistency across multiple source generators, and scaling to a larger backbone further extends the framework toward more general video refinement.

Preprint.

1 Introduction

Recent video generation models can synthesize visually compelling clips from text, images, or video conditions [1, 2, 3, 4, 5, 6, 7, 8]. However, their apparent ability to behave as world simulators often breaks down once the camera moves. Rigid structures may warp under viewpoint changes, parallax may become inconsistent, scene layout may drift across frames, and continuous camera motion may collapse into shot-like transitions. These failures are especially severe under large viewpoint changes, showing that current models often generate plausible frame sequences but fail to maintain a stable 3D scene layout. We argue that camera-motion-induced geometric distortion is largely a data problem. In web-scale video corpora, most visible motion comes from dynamic objects captured by static or weakly moving cameras, leaving only sparse supervision for viewpoint-induced appearance changes. As a result, camera motion and object motion are easily entangled: when asked to generate a moving-camera shot, a model may explain the visual change by deforming scene structure rather than maintaining a stable 3D layout.

To address this problem, existing works have largely followed two directions. The first makes the generator more geometry-aware by introducing 3D supervision or spatial priors during training [9, 10, 11, 12, 13]. However, as foundation models evolve rapidly and are often accessible only through checkpoints or inference APIs, modifying the source model is not always feasible. The second direction applies post-hoc video refinement [7, 14, 15]. These methods primarily target pixel-level restoration, such as super-resolution, denoising, and temporal smoothing, thereby improving appearance without explicitly correcting the underlying geometric distortion. Post-hoc geometry correction, which requires reasoning about 3D structure under camera motion, remains comparatively underexplored.

The key bottleneck is the training data. Training a geometry corrector requires paired supervision: a geometry-distorted video and a geometry-consistent target that follows the same camera trajectory. Such pairs are *nearly impossible* to collect in the real world. To address this challenge, we construct the *VG-Pairs* dataset using a paired-supervision data generation pipeline. Starting from real videos with diverse camera motion and annotated poses, we synthesize distorted counterparts using Wan 2.1-Fun-InP, a pretrained image-to-video inpainting model based on Wan [4]. Instead of conditioning only on the first and last frames, we adopt a trajectory-adaptive keyframe strategy that inserts additional keyframes at abrupt pose changes, enabling stable generation of shorter subsegments that are later stitched together. This process produces paired videos with identical semantics and similar camera motion, but different geometric fidelity.

Built on *VG-Pairs*, we propose Video Geometry Corrector (*VGC*), a video-to-video diffusion model that transforms geometry-distorted videos into structurally coherent outputs. *VGC* restores stable scene geometry without regenerating videos from scratch, preserving scene content and object appearance. To further disentangle camera motion from object motion, we incorporate optical flow as an auxiliary structural cue. Optical flow separates globally coherent camera-induced motion from local object dynamics, helping the model determine what to correct and what to preserve.

Although our main focus is correcting geometry in static scenes under camera motion, scaling *VGC* to larger-capacity models naturally extends its capability to dynamic videos containing human and object motion, moving toward general video refinement. This flexibility distinguishes *VGC* from prior geometry-focused approaches [16, 10], which typically focus on static environments. The main contributions of this work are:

- We identify a critical gap in post-hoc video refinement and attribute camera-induced geometric distortions to insufficient supervision for viewpoint-induced appearance changes.
- We address the paired-data bottleneck with *VG-Pairs*, a synthetic paired-supervision dataset built from real videos with camera motion and their trajectory-aware geometric distortions.
- We propose *VGC*, a post-hoc video-to-video geometry corrector augmented with optical flow cues, demonstrating strong structural restoration, cross-generator transferability, and scalability to dynamic videos.

2 Related Work

Geometry-consistent video generation. Recent video diffusion models such as Sora [1], CogVideoX [2], HunyuanVideo [3], and Wan [4] achieve strong visual quality, yet maintaining

geometric consistency under camera motion remains challenging. Most existing approaches address this issue at the generator or training level. Some methods introduce explicit geometric supervision during training. GeoVideo [9] incorporates depth prediction and cross-frame consistency losses. Geometry Forcing [10] aligns intermediate features with pretrained 3D foundation models [17], and Epipolar Geometry Improves Video Generation Models [11] introduces epipolar constraints during fine-tuning. Other approaches improve geometry through post-training alignment or reward optimization. VideoGPA [16] derives preference signals from reprojection consistency, VideoREPA [18] transfers physical reasoning from video foundation models, and VGGRPO [19] extends geometry-aware rewards into latent space for dynamic scenes. A third direction incorporates explicit 3D structure into the generation pipeline. World-consistent Video Diffusion [12] maintains a point-cloud memory for spatial consistency, while GEN3C [13] uses geometric priors for camera-controlled video generation. In contrast, *VGC* frames geometry correction as a post-hoc video-to-video task, avoiding modification of the source generator and enabling transfer across model families.

Video refinement and enhancement. A separate line of work improves generated or degraded videos through post-processing. Upscale-A-Video [7], SeedVR [14], and SeedVR2 [20] focus on restoration and super-resolution, progressing from multi-step diffusion upscaling to single-step arbitrary-resolution enhancement. Meanwhile, several works [8, 21] introduce new datasets to refine high-resolution video generation. VEnhancer [15] jointly improves spatial and temporal quality using a space-time controller on pretrained diffusion priors. Enhance-A-Video [22] enhances temporal consistency in DiT-based models without retraining. ReVision [23] refines coarse videos using 3D object-centric representations and physical motion priors. Despite improving visual quality, temporal smoothness, and motion plausibility, these methods largely overlook post-hoc correction of scene-level geometric consistency under camera motion. *VGC* differs by targeting geometry correction, using paired supervision constructed from real videos to repair structural inconsistency while preserving scene content and motion.

3 Method

Our goal is to restore geometric consistency in generated videos under camera motion. We attribute these failures to insufficient supervision for viewpoint-induced appearance change in existing video data. Instead of modifying each video generation model to learn stronger geometric priors, we formulate geometry correction as a post-hoc video-to-video task and train a separate corrector with paired supervision. This decouples geometry correction from video generator design, making our method applicable across different model families.

Our framework consists of two stages. First, we build *VG-Pairs* by synthesizing paired distorted inputs and geometry-consistent targets from real videos with diverse camera motion and pose annotations. Second, we train a video-to-video diffusion model, *VGC*, on these pairs to correct geometry. Optical flow extracted from the distorted input serves as an auxiliary cue for disentangling camera motion from scene dynamics.

3.1 Paired Supervision from Videos with Camera Motion

The central design choice is to use videos with diverse camera motion not merely as additional training data for a generator, but as a source of correction pairs for a video-to-video task. Hence, we construct *VG-Pairs*, a dataset of geometry-distorted videos paired with geometry-consistent supervision, using the following data generation pipeline. Let $\mathbf{x} = (x_1, \ldots, x_T)$ denote a real video clip with camera poses $\mathbf{p} = (p_1, \ldots, p_T)$, where each pose is represented as $p_i = (\mathbf{t}_i, \mathbf{q}_i)$ with camera position $\mathbf{t}_i \in \mathbb{R}^3$ and rotation $\mathbf{q}_i \in S^3$. We treat $\mathbf{x}$ as the geometry-consistent target and synthesize a distorted counterpart, $\mathbf{y}$, using Wan 2.1-Fun-InP, a pretrained image-to-video inpainting model based on Wan [4]. Given a start frame and an end frame, this model generates the intermediate video that connects them.

To make $\mathbf{y}$ useful for supervision, it should preserve the scene identity and approximately follow the camera trajectory of $\mathbf{x}$ while still exhibiting geometric errors. We therefore select a trajectory-adaptive set of keyframes $K = \{k_1, \ldots, k_{m+1}\}$ with $1 = k_1 < \cdots < k_{m+1} = T$, and generate one video segment between each consecutive keyframe pair. Figure 2 illustrates the motivation: for clips with large, nonlinear camera motion, conditioning only on the first and last frames can lead the inpainting model to produce a shortcut trajectory that deviates from the ground-truth path, whereas adaptive keyframe selection yields distorted videos that remain close to the target trajectory.

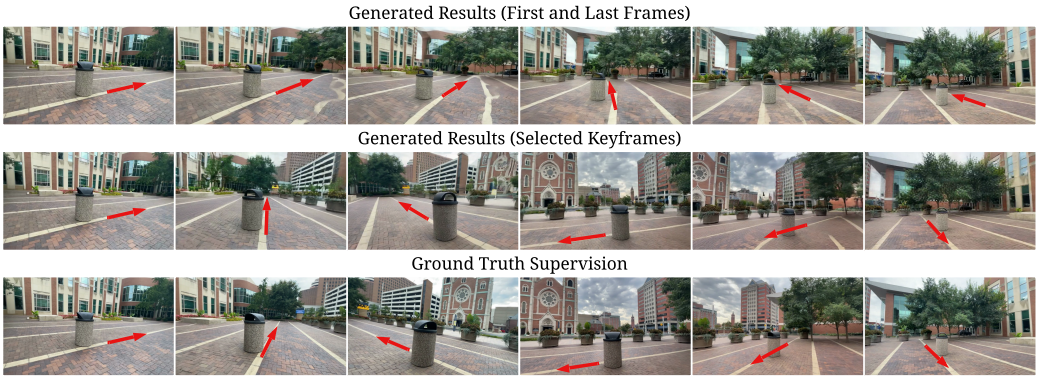


Figure 2: **Effect of adaptive keyframe selection.** The red arrows indicate a constant 3D direction while the camera is moving. For clips with large, nonlinear camera motion, conditioning the image-to-video model only on the first and last frames often produces a shortcut trajectory that deviates substantially from the ground-truth path (top; about 45° rotation versus about 225° in the ground truth). Our adaptive keyframe selection inserts intermediate keyframes, generates shorter segments conditioned on their boundary keyframes, and stitches them together. This produces geometry-distorted videos that remain close to the target camera trajectory and are therefore suitable as paired training data (middle), compared with the ground-truth supervision video (bottom).

We first estimate the total amount of camera motion in the clip. For each adjacent frame pair, we measure rotational change by the geodesic distance between quaternions, $\Delta\phi_i = 2\arccos(|\langle \mathbf{q}_i, \mathbf{q}_{i+1}\rangle|)$, and translational change by the Euclidean displacement $\|\mathbf{t}_{i+1} - \mathbf{t}_i\|_2$. Summing these local changes over the full clip gives the total rotation and translation:

$$R_{\text{tot}} = \sum_{i=1}^{T-1} \Delta\phi_i, \qquad D_{\text{tot}} = \sum_{i=1}^{T-1} \|\mathbf{t}_{i+1} - \mathbf{t}_i\|_2. \tag{1}$$

We then choose how many trajectory segments are needed so that each segment covers only a manageable amount of motion. Concretely, R_{seg} and D_{seg} denote the target rotation and translation budget for a single segment. We compare the total motion of the clip against these budgets along both axes and let the larger normalized motion determine the number of segments, capped at $M_{\max}$:

$$m = \min\left(M_{\max}, \max\left(1, \left\lceil \max\left(\frac{R_{\text{tot}}}{R_{\text{seg}}}, \frac{D_{\text{tot}}}{D_{\text{seg}}}\right)\right\rceil\right)\right). \tag{2}$$

If necessary, we further reduce m so that each segment contains at least $L_{\min}$ frames. As a result, short or nearly linear trajectories are handled with a single segment, whereas clips with larger accumulated motion are divided into shorter intervals that are easier for the inpainting model to follow.

Once m is fixed, we place the $m-1$ interior keyframes at moments of peak trajectory inflection, the exact points where the generator is most prone to geometric drift. To capture such changes, we compute the local translation and rotation increments, $\mathbf{v}_i = \mathbf{t}_{i+1} - \mathbf{t}_i$ and $\boldsymbol{\omega}_i = \log(\mathbf{q}_{i+1}\mathbf{q}_i^{-1})$, and assign each interior frame a motion-change score:

$$s(i) = \theta(\mathbf{v}_{i-1}, \mathbf{v}_i) + \alpha\,\theta(\boldsymbol{\omega}_{i-1}, \boldsymbol{\omega}_i), \tag{3}$$

where $\theta(\cdot,\cdot)$ denotes the angle between two vectors, and each term is suppressed when the corresponding motion magnitude is below a small threshold. A large score indicates a sharp change in translation direction, rotation direction, or both. We select up to $m-1$ interior keyframes by greedy non-maximum suppression over $s(i)$, enforcing a minimum temporal separation of $L_{\min}$. If fewer than $m-1$ reliable peaks are selected, we place the remaining keyframes by approximately uniform partition of the largest uncovered gaps.

Given the final keyframe set K, we generate each segment independently by conditioning the inpainting model on the boundary frames x_{k_j} and $x_{k_{j+1}}$. Concatenating the resulting subclips and removing duplicated boundary frames yields the distorted counterpart $\mathbf{y}$. This adaptive construction keeps the synthesized video close enough to the target that the learning problem remains correction rather than regeneration, while still inducing the geometric failures of interest, such as structural warping, inconsistent parallax, depth-order violations, and shot-like transitions.

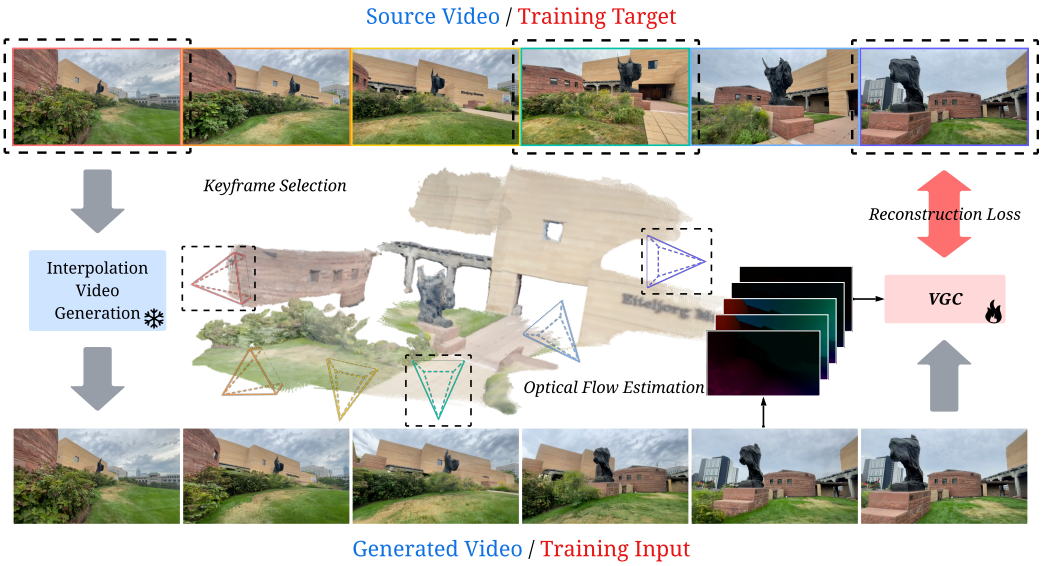


Figure 3: **Overview of Video Geometry Corrector (*VGC*).** Starting from a real video with diverse camera motion, we select trajectory-adaptive keyframes and use them to condition an image-to-video inpainting model, which produces a geometry-distorted video that serves as the training input. We then train *VGC* to map this distorted input to the geometry-consistent target video with a reconstruction loss. Optical flow is computed from the distorted video and used as an additional conditioning signal for *VGC*, providing motion cues that help disentangle camera motion from scene dynamics. *The 3D reconstruction is shown here only for visualization and is not used in the pipeline.*

3.2 Motion Disentanglement with Optical Flow

Geometry distortions under camera motion arise in part because current video training data contains rich human and object motion, but relatively sparse supervision for viewpoint-induced appearance change. As a result, video models can confuse camera motion with scene dynamics, and may explain viewpoint change by deforming scene structure instead of preserving a stable 3D layout. Geometry correction is therefore not only a restoration problem, but also a motion disentanglement problem.

Optical flow provides a natural auxiliary representation for this purpose. Unlike RGB frames, which entangle appearance and motion, optical flow explicitly represents motion through its location, direction, and magnitude. Camera motion tends to induce a globally coherent motion field, whereas object motion appears as local deviations from that field. We therefore compute forward optical flow between consecutive frames of the distorted input video and use it as an additional conditioning signal for the correction model. Because the flow is computed from the distorted input rather than the target, the same signal is available during both training and inference. Optical flow does not supervise the model directly; instead, it provides motion structure that helps distinguish what should be corrected from what should be preserved.

3.3 Video Geometry Corrector

Figure 3 summarizes the overall data generation and training pipeline. *VGC* is built on a pretrained video diffusion backbone [4] and trained to map a distorted input video $\mathbf{y}$ to its geometry-consistent target $\mathbf{x}$. Let $E(\cdot)$ denote the video encoder, and let $\mathbf{z} = E(\mathbf{x})$ be the latent representation of the target video. During training, we sample a diffusion timestep t and Gaussian noise ϵ, and construct the noisy latent $\mathbf{z}_t$ in the standard way. The denoising network is trained to predict ϵ conditioned on the distorted input video, the clip text prompt, and the optical flow maps:

$$\mathcal{L} = \mathbb{E}_{t,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2\right], \tag{4}$$

where $\mathbf{c}$ denotes the distorted-video latent, the text condition, and the optical flow condition. We inject the video and flow conditions through channel concatenation, while the text prompt is provided through cross-attention. The model is thus trained to preserve scene content and camera intent while denoising toward a geometry-consistent target. At inference, *VGC* takes a distorted input video, computes optical flow from it, and performs conditional denoising to produce a geometry-corrected

output. Because the model is trained as a separate video-to-video corrector, it can refine videos from different generators without modifying the source model itself.

4 Experiments

4.1 Experimental Setup

Implementation details. We construct *VG-Pairs* from two publicly available datasets, DL3DV [24] and SpatialVID [25]. We select 9K videos from DL3DV, reserve the first 1K for evaluation, and filter 63K videos from SpatialVID that contain large camera motion. To construct paired supervision, we use Wan 2.1-Fun-InP to synthesize geometry-distorted videos from trajectory-adaptive keyframes. Given a start frame and an end frame, Wan 2.1-Fun-InP generates the intermediate video connecting them. For clips with complex camera motion, we apply this model independently to multiple keyframe-bounded subsegments and stitch the generated segments together to form the distorted training input. Because Wan's VAE temporally compresses the video, keyframes are selected so that each generated subsegment has a frame count compatible with the temporal compression schedule. We use Wan 2.1-1.3B [4] as the backbone of *VGC*. The model is trained for 10K steps on 32 H100 GPUs with AdamW, using a learning rate of 1×10^{-5}, weight decay 0.01, and a global batch size of 64. Unless otherwise specified, all training videos contain 81 frames at a resolution of 832×480. Optical flow is estimated with SEA-RAFT [26] and used as an additional conditioning signal. We use $M_{max} = 4$, $R_{seg} = 120°$, $D_{seg} = 10$ m, $L_{min} = 15$, and $\alpha = 1$.

Evaluation datasets and metrics. To evaluate *VGC*'s performance, we study two scenarios in which camera motion is introduced. In the first setting, camera motion is specified by camera poses. Following Geometry Forcing [10], we evaluate on RealEstate10K [27] using 100 randomly sampled test videos and DFoT [28] as the source generator. We compare against Geometry Forcing and Geometry Forcing + REPA [29], and report PSNR, SSIM, LPIPS [30], and revisit error (RVE). Following [10], RVE measures long-range consistency under a full camera rotation by comparing the generated view at the end of the trajectory with the corresponding reference view. In the second setting, camera motion is induced by the text prompt. We use three source generators in this setting: CogVideoX-5B I2V, CogVideoX-5B T2V, and CogVideoX 1.5-5B T2V. We randomly sample 200 prompts from the held-out DL3DV test set. Following VideoGPA [16], we reconstruct each generated video independently using Depth Anything 3 [31] and compare its original frames with frames reprojected from the reconstructed 3D geometry to compute PSNR, SSIM, LPIPS [30], MSE, and 3D consistency score (3DCS). Unlike the pose-conditioned setting, these metrics do not require ground-truth video frames. In both settings, *VGC* is applied as a post-hoc corrector to the outputs of the source generator, demonstrating its flexibility across different video model families.

Table 1: **Quantitative results on RealEstate10K [27].** *VGC* achieves the strongest PSNR, SSIM, and revisit error (RVE) among the direct baselines. When applied as a post-hoc corrector to Geometry Forcing (GF) [10] and GF + REPA [10, 29], it further improves their performance, showing that it can be flexibly combined with geometry-aware generators. Comparing *VGC* with and without optical flow shows that flow yields modest gains in image quality while maintaining similar long-range geometric consistency.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	RVE↓
DFoT [28]	13.61	0.4006	0.4487	184.2
GF [10]	14.76	0.4484	**0.4127**	181.0
GF [10] + REPA [29]	14.16	0.4217	0.4218	171.8
DFoT + *VGC* (w/o Flow)	14.76	0.4509	0.4433	169.4
DFoT + *VGC* (Ours)	**14.87**	**0.4545**	0.4395	**168.9**
GF + *VGC*	15.50 (+0.74)	0.4715 (+0.0232)	0.4150 (+0.0022)	158.8 (-22.2)
GF + REPA + *VGC*	15.25 (+1.09)	0.4675 (+0.0458)	0.4226 (+0.0008)	160.9 (-10.9)



DFoT | Geometry Forcing | Geometry Forcing + REPA | DFoT + *VGC* (Ours)

Figure 4: **Visual results on RealEstate10K [27].** DFoT [28] and GF + REPA [10, 29] produce inconsistent lighting across frames, while GF [10] generates an incomplete sofa structure (red box). *VGC* restores more consistent lighting and more coherent scene geometry.

4.2 Experimental Results

Results on RealEstate10K. Table 1 shows that *VGC* improves geometric consistency in the pose-conditioned setting and achieves stronger PSNR, SSIM, and RVE than the DFoT baseline [28]. When applied as a post-hoc corrector to the outputs of Geometry Forcing [10] and Geometry Forcing + REPA [10, 29], *VGC* further improves their results, demonstrating that it can be flexibly combined with geometry-aware generators. Figure 4 provides qualitative examples of the improved temporal

Table 2: **Quantitative results on DL3DV [24].** We evaluate *VGC* on three CogVideoX [2] source generators in both text-to-video and image-to-video settings. *VGC* consistently improves PSNR, SSIM, and MSE over the corresponding baselines, and further improves VideoGPA [16] when applied as a post-hoc corrector. These results show that *VGC* transfers across different generation settings and can be flexibly combined with other geometry-aware refinement methods.

Method	*CogVideoX-5B T2V*					*CogVideoX 1.5-5B T2V*					*CogVideoX-5B I2V*				
	PSNR↑	SSIM↑	LPIPS↓	MSE↓	3DCS↓	PSNR↑	SSIM↑	LPIPS↓	MSE↓	3DCS↓	PSNR↑	SSIM↑	LPIPS↓	MSE↓	3DCS↓
Baseline	15.6668	0.5389	0.5285	0.0333	0.8617	15.9062	0.4644	0.5449	0.0289	0.8342	17.3189	0.5543	0.5760	0.0218	0.7944
Baseline + *VGC*	16.6940 (+1.0272)	0.5519 (+0.0130)	0.5733 (+0.0448)	0.0264 (-0.0069)	0.8376 (-0.0241)	17.7342 (+1.8280)	0.5284 (+0.0640)	0.5936 (+0.0487)	0.0192 (-0.0097)	0.7859 (-0.0483)	17.9989 (+0.6800)	0.5456 (-0.0087)	0.6284 (+0.0524)	0.0184 (-0.0034)	0.8121 (+0.0177)
VideoGPA [16]	14.6318	0.5210	0.5065	0.0428	0.9346	14.4002	0.4175	0.5531	0.0408	0.9614	15.7203	0.5240	0.5749	0.0320	0.8952
VideoGPA [16] + *VGC*	15.9092 (+1.2774)	0.5413 (+0.0203)	0.5546 (+0.0481)	0.0319 (-0.0109)	0.8737 (-0.0609)	16.8430 (+2.4428)	0.5039 (+0.0864)	0.5907 (+0.0376)	0.0235 (-0.0173)	0.8258 (-0.1356)	17.3053 (+1.5850)	0.5379 (+0.0139)	0.6216 (+0.0467)	0.0225 (-0.0095)	0.8463 (-0.0489)

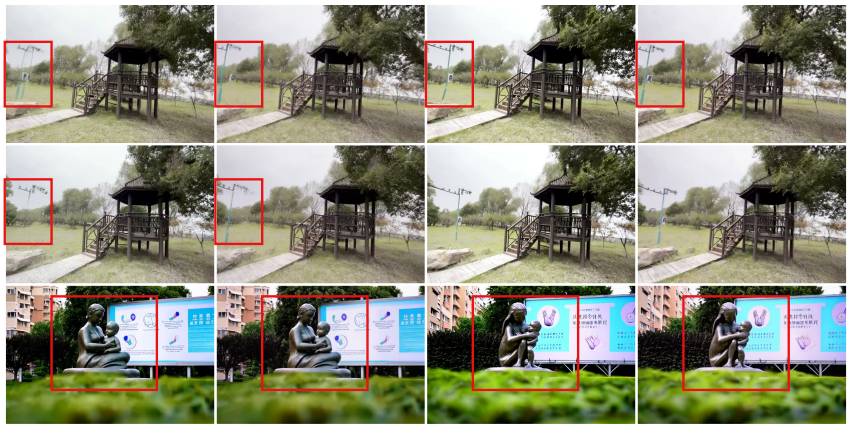


Figure 5: **Visual results on DL3DV [24].** The first two rows show image-to-video results. In the baseline CogVideoX [2] outputs, the streetlight is duplicated, while VideoGPA [16] produces a distorted and tilted streetlight (red boxes). *VGC* corrects these geometric errors and restores a more plausible structure. The last row shows a text-to-video example, where *VGC* better preserves the original appearance of the sculpture, while VideoGPA changes its structure and other scene details.

consistency and scene structure. Overall, these results support our claim that *VGC* improves both local visual quality and cross-frame geometric consistency under camera motion.

We also conduct an ablation study in this setting to evaluate the effect of optical flow as an auxiliary conditioning signal. Comparing the full model with the variant without flow in Table 1, we find that optical flow improves PSNR, SSIM, and RVE. These results suggest that flow helps improve local structural consistency, and that *VGC*'s main gains come from the paired supervision itself.

Results on DL3DV. Table 2 shows that *VGC* improves reconstruction quality and geometric consistency across all three CogVideoX source generators in both text-to-video and image-to-video settings. Compared with the original baselines, *VGC* consistently improves PSNR, SSIM, and MSE, indicating better frame fidelity under prompt-driven camera motion. When applied on top of VideoGPA [16], *VGC* further improves its results, showing that the proposed corrector can be flexibly combined with geometry-aware refinement methods. Figure 5 provides qualitative examples in which *VGC* corrects geometric distortions while better preserving scene appearance.

4.3 Extension to General Video Refinement

We further investigate whether the proposed correction framework can extend beyond geometry correction in static scenes. To this end, we train a larger version of *VGC* based on Wan 2.2-14B and evaluate it in a more general video refinement setting. With increased model capacity, *VGC* not only corrects geometric distortions under camera motion, but also improves dynamic videos by recovering more plausible human and object geometry. As shown in Figure 6, the smaller 1.3B version does not exhibit the same ability to recover dynamic motion. Because the 14B model takes 480p videos as input and produces 720p outputs, it also performs super-resolution as part of the refinement process.

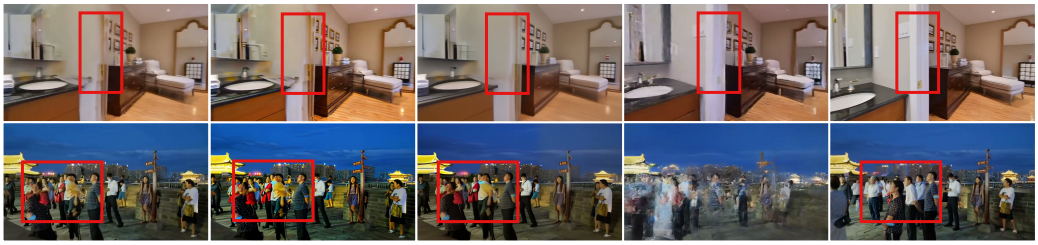


Figure 6: **Comparison of model sizes.** In static scenes (top), both *VGC* 1.3B and 14B correct geometric distortions produced by the baseline model that SeedVR2 [20] and VEnhancer [15] fail to restore (red boxes). The 14B model further improves the geometry of moving humans and objects in dynamic scenes (bottom).

Table 3: **General video refinement results on SpatialVID [25].** We compare *VGC* with general video refinement and enhancement methods [20, 15] on dynamic videos generated by CogVideoX [2], HunyuanVideo [3], and Wan [4]. Although *VGC* is designed for geometry correction under camera motion rather than generic enhancement, it achieves competitive performance across multiple perceptual and temporal metrics, and is particularly strong in dynamic degree, imaging quality, and DOVER [32]. These results suggest that the proposed correction framework extends beyond static-scene geometry and remains effective in more general settings.

Method	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	DOVER [32]
CogVideoX-I2V [2]								
Baseline	0.9480	0.9455	0.9742	0.9918	0.30	0.5095	69.6945	9.54
SeedVR2 [20]	0.9447	0.9476	0.9603	0.9829	0.30	**0.5313**	71.1767	8.31
VEnhancer v2 [15]	**0.9506**	**0.9602**	0.9749	**0.9920**	0.31	0.5237	69.3742	10.31
VGC (Ours)	0.9483	0.9450	**0.9772**	0.9891	**0.66**	0.5011	**72.7654**	**10.50**
HunyuanVideo 1.5-I2V [3]								
Baseline	0.8970	0.9231	0.9592	0.9829	**1.00**	0.5133	73.7167	**10.10**
SeedVR2 [20]	0.8967	0.9167	0.9415	0.9714	**1.00**	**0.5539**	72.2762	7.89
VEnhancer v2 [15]	**0.9061**	**0.9294**	**0.9596**	**0.9860**	0.99	0.5301	71.9688	9.85
VGC (Ours)	0.8995	0.9282	0.9482	0.9799	0.96	0.4849	**75.1367**	9.30
Wan 2.1-Fun-InP [4]								
Baseline	0.9247	0.9239	**0.9712**	**0.9914**	0.87	0.5043	69.5165	10.87
SeedVR2 [20]	0.9217	0.9231	0.9547	0.9809	0.87	**0.5431**	**72.9463**	8.23
VEnhancer v2 [15]	0.9242	**0.9371**	0.9702	0.9909	0.85	0.5254	70.6322	10.78
VGC (Ours)	**0.9268**	0.9281	0.9641	0.9907	**0.89**	0.5141	72.6499	**11.47**

For evaluation, we randomly sample 100 held-out videos from SpatialVID [25] and compare against general video refinement methods, including SeedVR2 [20] and VEnhancer v2 [15]. To demonstrate the flexibility of *VGC*, we apply it to videos generated by three different source models: CogVideoX [2], HunyuanVideo 1.5 [3], and Wan 2.1 [4]. CogVideoX and HunyuanVideo 1.5 are evaluated in the image-to-video setting, while Wan 2.1 is evaluated in the inpainting setting. Since these videos contain dynamic content and do not admit a geometry-consistent reconstruction target, we evaluate them with no-reference video-quality metrics rather than reconstruction metrics. Specifically, we report VBench [33] metrics together with DOVER [32], a holistic no-reference video-quality metric that summarizes technical and aesthetic quality.

The quantitative results in Table 3 show that *VGC* achieves performance comparable to strong video refinement baselines across a range of perceptual and temporal metrics, despite not being explicitly designed for generic enhancement. In particular, it remains consistently competitive in subject consistency, background consistency, and motion quality, while achieving strong results in dynamic degree, imaging quality, and DOVER. These results suggest that, when scaled to larger backbones, the proposed correction framework naturally generalizes to broader video refinement settings.

Figure 7 provides qualitative examples. *VGC* corrects implausible geometry changes, recovers more plausible camera motion, restores distorted structural details, and better preserves human identity and motion continuity. It also improves local detail through super-resolution while remaining faithful to the source video.

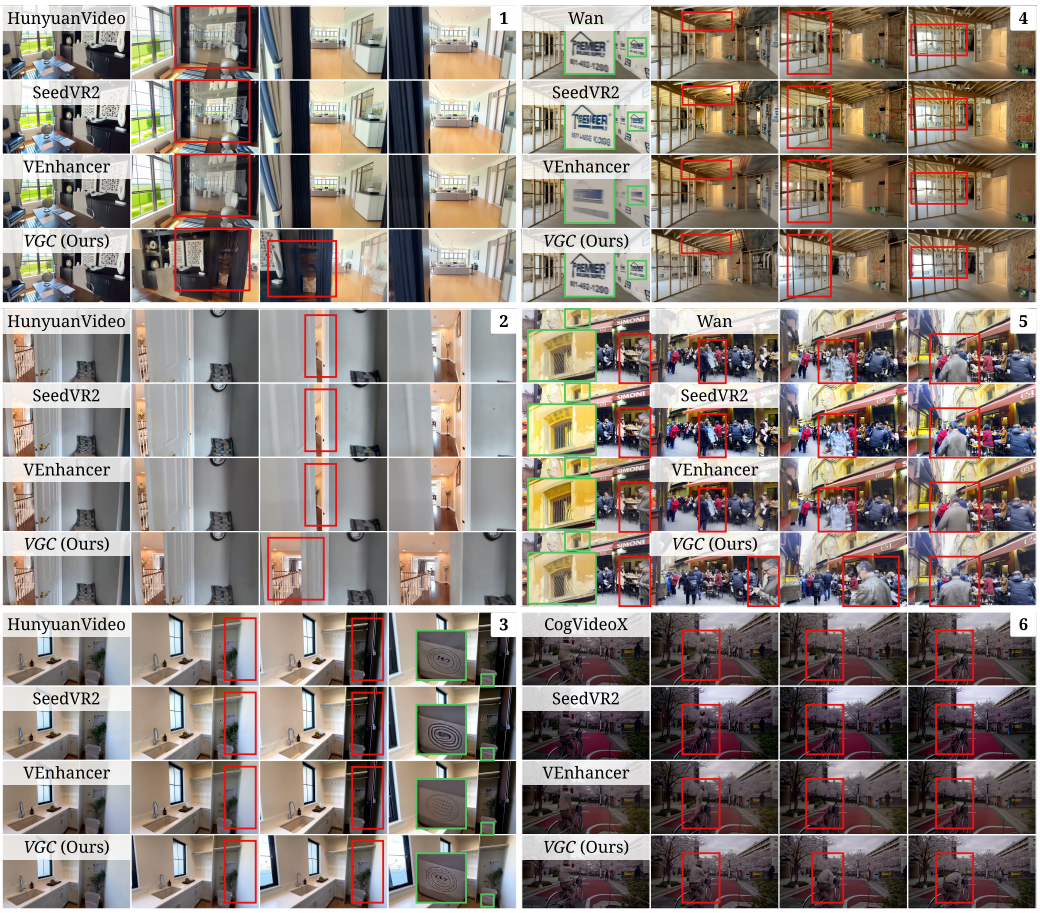


Figure 7: **General video refinement results.** *VGC* corrects several geometric failures produced by the baseline models. In scene 1, HunyuanVideo [3] changes the cabinet into a glass door, while *VGC* restores the correct scene structure and camera motion, placing the door on the right side of the cabinet. In scene 2, HunyuanVideo generates a camera trajectory that passes through the wrong side of the doorway, whereas *VGC* recovers a plausible path. In scenes 3 and 4, *VGC* restores a consistent appearance and recovers fine geometric structure, such as the wooden supports. In the dynamic scenes, *VGC* better preserves identity and motion continuity: in scene 5, it maintains the same man in the leather jacket across frames and recovers a plausible motion trajectory, whereas the Wan baseline fails to infer coherent human motion; in scene 6, it restores the missing cyclist produced by CogVideoX [2]. The green boxes further show that *VGC* enhances local detail while remaining more faithful to the source video than SeedVR2 [20] and VEnhancer [15]. *Please zoom in for details.*

5 Conclusion

We presented Video Geometry Corrector (*VGC*), a post-hoc video-to-video model for correcting geometric distortions in generated videos under camera motion. We argued that these failures are largely a data problem: current video training data provides limited supervision for viewpoint-induced appearance change, causing camera motion to be entangled with scene dynamics. To address this, we constructed *VG-Pairs* from real videos with diverse camera motion and trained *VGC* to map geometry-distorted videos to geometry-consistent targets. We further incorporated optical flow as an auxiliary cue to help distinguish what should be corrected from what should be preserved. Experiments on pose-conditioned and prompt-driven benchmarks show that *VGC* improves geometric consistency across different source generators and can be flexibly combined with existing geometry-aware methods. With increased model capacity, the same correction framework also extends beyond static-scene geometry toward more general video refinement.

Acknowledgments

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project) and SONY.

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*, 2024. 2

[2] 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. In *ICLR*, 2025. 2, 7, 8, 9

[3] Tencent Hunyuan Foundation Model Team. Hunyuanvideo 1.5 technical report, 2025. 2, 8, 9

[4] 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, 3, 5, 6, 8

[5] Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026. 2

[6] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Kai Chen, and Yunhai Tong. Motionbooth: Motion-aware customized text-to-video generation. In *NeurIPS*, 2024. 2

[7] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-A-Video: Temporal-consistent diffusion model for real-world video super-resolution. In *CVPR*, 2024. 2, 3

[8] Jiangning Zhang, Junwei Zhu, Teng Hu, Yabiao Wang, Donghao Luo, Weijian Cao, Zhenye Gan, Xiaobin Hu, Zhucun Xue, Xiangtai Li, Chengjie Wang, and Yong Liu. Transform trained transformer: Accelerating naive 4k video generation over 10 times. In *ICML*, 2026. 2, 3

[9] Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. Geovideo: Introducing geometric regularization into video generation model. *arXiv preprint arXiv:2512.03453*, 2025. 2, 3

[10] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry Forcing: Marrying video diffusion and 3D representation for consistent world modeling. In *ICLR*, 2026. 2, 3, 6

[11] Orest Kupyn, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models. *arXiv preprint arXiv:2510.21615*, 2025. 2, 3

[12] Qihang Zhang, Shuangfei Zhai, Miguel Ángel Bautista, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *CVPR*, 2025. 2, 3

[13] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *CVPR*, 2025. 2, 3

[14] Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In *CVPR*, 2025. 2, 3

[15] Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. VEnhancer: Generative space-time enhancement for video generation. *arXiv preprint arXiv:2407.07667*, 2024. 2, 3, 8, 9

[16] Hongyang Du, Junjie Ye, Xiaoyan Cong, Runhao Li, Jingcheng Ni, Aman Agarwal, Zeqi Zhou, Zekun Li, Randall Balestriero, and Yue Wang. VideoGPA: Distilling geometry priors for 3D-consistent video generation. *arXiv preprint arXiv:2601.23286*, 2026. 2, 3, 6, 7

[17] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025. 3

[18] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. VideoREPA: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025. 3

[19] Zhaochong An, Orest Kupyn, Théo Uscidda, Andrea Colaco, Karan Ahuja, Serge J. Belongie, Mar González-Franco, and Marta Tintore Gazulla. VGGRPO: towards world-consistent video generation with 4d latent reward. In *ECCV*, 2026. 3

[20] Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, et al. Seedvr2: One-step video restoration via diffusion adversarial post-training. *arXiv preprint arXiv:2506.05301*, 2025. 3, 8, 9

[21] Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. In *NeurIPS*, 2025. 3

[22] Yang Luo, Xuanlei Zhao, Mengzhao Chen, Kaipeng Zhang, Wenqi Shao, Kai Wang, Zhangyang Wang, and Yang You. Enhance-A-Video: Better generated video for free. *arXiv preprint arXiv:2502.07508*, 2025. 3

[23] Qihao Liu, Ju He, Qihang Yu, Liang-Chieh Chen, and Alan Yuille. ReVision: Refining video diffusion with explicit 3D motion modeling. *arXiv preprint arXiv:2504.21855*, 2025. 3

[24] 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*, 2024. 6, 7

[25] Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025. 6, 8

[26] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *ECCV*, 2024. 6

[27] 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. 6

[28] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *ICML*, 2025. 6

[29] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *ICLR*, 2025. 6

[30] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[31] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *ICLR*, 2026. 6

[32] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023. 8

[33] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. 8