CANOCAM: GIMBAL AND INTRINSICS CONTROL FOR VIDEO GENERATION VIA CANONICAL VIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Camera-controlled video generation remains unreliable under extreme pitch and roll, wide fields of view, and strong lens distortion. These camera configurations are rare in training data, making their visual effects difficult to learn jointly with scene dynamics. We present **CanoCam**, which reduces this learning burden through canonical video generation and geometric reprojection. Our approach exploits a geometric distinction: translation induces depth-dependent parallax and visibility changes, whereas rotation and changes in intrinsics admit a depth-independent ray mapping at a fixed camera center. CanoCam first generates canonical videos under simplified camera conditions while preserving the requested camera-center trajectory. We consider two canonical representations: level perspective views that retain yaw and panoramic views that provide full angular coverage. Reprojection then applies the remaining rotation and target intrinsics. For both representations, a target-camera-conditioned video model completes missing regions and refines the reprojected content. Experiments on SpatialVID-Extreme, Prompt Entanglement, and PanShot show improved camera control over strong baselines across extreme orientations and diverse camera models, including both pinhole and fisheye cameras. Controlled ablations demonstrate the importance of canonical generation and target-camera-conditioned refinement for accurate camera control. The code will be made publicly available upon acceptance.

1 INTRODUCTION

Recent video diffusion models generate increasingly compelling visual content (Wan et al., 2025), opening up new possibilities for visual storytelling. Yet fine-grained control over camera configuration remains limited. Beyond camera translation, pitch, roll, field of view, and lens distortion fundamentally shape how a scene appears. An upward tilt, a sideways roll, or a fisheye lens can produce dramatically different views of the same scene. Reliably realizing such camera configurations, especially under extreme orientations and wide-angle lenses, remains challenging for controllable video generation.

Existing camera-conditioned approaches inject camera parameters into the denoising process (He et al., 2025a; Bahmani et al., 2025a;b). UCPE (Zhang et al., 2026a) jointly represents camera pose, intrinsics, lens distortion, and gravity-referenced orientation, while GimbalDiffusion (Fortier-Chouinard et al., 2026) improves absolute orientation control through panoramic training data and null-pitch conditioning. Despite these advances, camera control remains tightly coupled with video generation: the model must learn camera-induced appearance changes while simultaneously synthesizing scene content, motion, and temporal dynamics. This coupling becomes particularly difficult for camera configurations that are sparsely represented in training data.

Our key observation is that not all camera changes need to be learned by the video model. Camera translation changes visibility and induces depth-dependent parallax, whereas orientation and intrinsics can be transformed through a depth-independent ray mapping when the camera center is fixed. InfCam (Kim et al., 2025) exploits related geometry to warp an input video with infinite homographies while learning residual parallax. We instead use this distinction to redesign the generation problem itself. We introduce **CanoCam**, which generates videos along the requested camera-center trajectory under a restricted canonical camera family, then applies the remaining orientation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

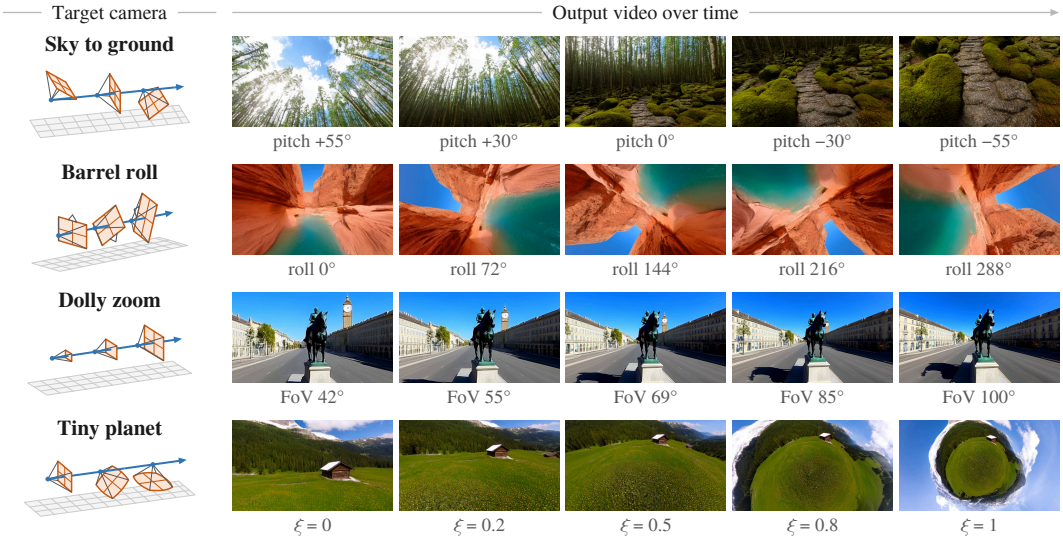

Figure 1: **Camera control by canonical generation and exact reprojection.** Each row is one generated video whose camera center follows the target path (blue); the target rotation and intrinsics (orange) are then applied by exact ray reprojection, here without refinement. The dolly zoom uses the level-view generator of CanoCam-P and the other rows the panoramic generator of CanoCam-E. Frames run left to right over time and are labeled with the parameter that changes along the row.

and projection changes geometrically. In this decomposition, generation handles scene dynamics and translation-induced parallax, while reprojection handles camera transformations that are known from geometry.

We instantiate this idea with two canonical representations. A *level-perspective* source preserves translation and yaw while fixing pitch, roll, and pinhole intrinsics. A *panoramic* source preserves translation while providing full angular coverage, allowing all target rotation to be deferred to reprojection. The two representations trade angular coverage for spatial resolution: perspective views concentrate detail within a limited frustum, while panoramas distribute resolution over the sphere. Reprojection may expose unseen regions or introduce artifacts, so we use a target-camera-conditioned refinement model to complete missing content and refine the projected video in the requested view.

Our contributions are threefold:

- We introduce **CanoCam**, a camera-controlled video generation framework that reduces the learning burden through canonical generation and geometric reprojection.
- We develop two alternative pipelines based on level-perspective and panoramic generation, offering different trade-offs between spatial resolution and angular coverage. Both incorporate target-camera-conditioned refinement to complete missing regions and refine the reprojected video.
- We demonstrate improved camera control over strong baselines on challenging benchmarks. Controlled experiments validate canonical generation and target-camera conditioning of the refinement model, and show that the decomposition also applies to a second video backbone, where it reduces GeoCalib orientation errors and relative translation errors.

2 Related Work

2.1 Video Generative Models

Video generation combines diffusion modeling (Ho et al., 2022b;a; Blattmann et al., 2023; Wang et al., 2023) with scalable transformers (Hong et al., 2023; Yang et al., 2025b; Wan et al., 2025).

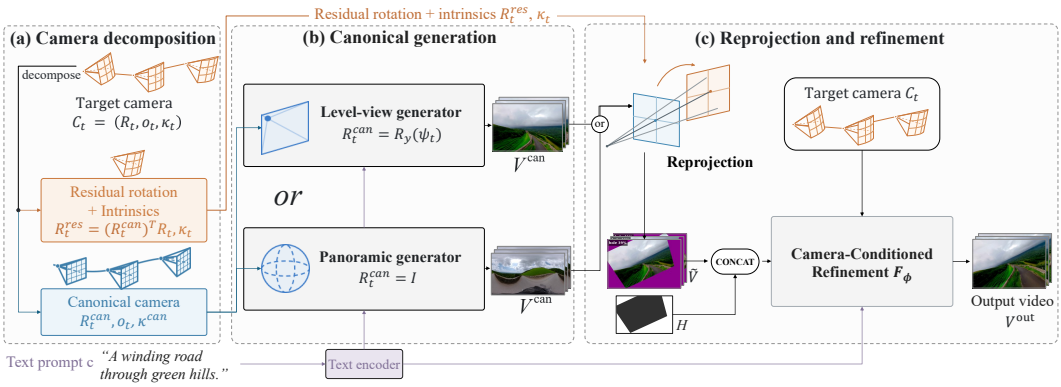

Figure 2: **Overview of CanoCam.** (a) Generation preserves the camera-center trajectory $\mathbf{o}_t$; reprojection applies the remaining rotation $R_t^{\text{res}} = (R_t^{\text{can}})^\top R_t$ and intrinsics κ_t. (b) The generator produces level perspective views retaining yaw (CanoCam-P) or panoramas (CanoCam-E). (c) Reprojection maps these videos to the target cameras. A camera-conditioned model completes missing regions (magenta) and refines the projected content. Both learned stages receive the text prompt.

Depth, flow, and physics priors guide structure and motion (Hong et al., 2022; Liang et al., 2024; Wang et al., 2025a). VideoPainter (Bian et al., 2025), HomoGen (Ding et al., 2025), and VACE (Jiang et al., 2025) use observed context for video completion and editing. Panoramic models adapt video priors to spherical imagery (Wang et al., 2024a; Xia et al., 2025) or extend perspective videos to full coverage (Tan et al., 2025; Luo et al., 2025).

2.2 Camera-Conditioned Video Generation

Camera-conditioned models encode poses or rays (Wang et al., 2024b; He et al., 2025a; Bahmani et al., 2025b;a). Recent methods support dynamic scene exploration (He et al., 2025b), video rerendering (Bai et al., 2025a), multimodal generation (Yang et al., 2025a), and synchronized views (Bai et al., 2025b). ViewCrafter and RealCam-I2V use reconstructed geometry (Yu et al., 2025; Li et al., 2025b), whereas our same-center reprojection requires no scene depth.

Perspective Fields (Jin et al., 2023) encode per-pixel camera attributes, while PRoPE (Li et al., 2025a) encodes relative camera frustums. UCPE (Zhang et al., 2026a) supports pose, intrinsics, lens distortion, and gravity-aware orientation. CanoCam complements camera conditioning by reducing the camera variation assigned to generation and applying the remaining transformations geometrically.

3 Method

Given a text prompt c and target cameras $\mathcal{C} = \{(R_t, \mathbf{o}_t, \kappa_t)\}_{t=1}^T$, we aim to generate a video consistent with the text and the specified camera poses and intrinsics. Here, $R_t \in \mathrm{SO}(3)$ and $\mathbf{o}_t \in \mathbb{R}^3$ denote the camera-to-world rotation and translation, respectively, in a gravity-aligned world frame. The projection mapping $\kappa_t(\cdot) : \mathbb{R}^3 \to \mathbb{R}^3$ encodes the camera intrinsics, including field of view (FoV) and lens distortion.

Learning this mapping directly is challenging for camera configurations that are poorly represented in training data, such as extreme tilt, wide FoV, and strong lens distortion. CanoCam addresses this challenge by exploiting a geometric distinction: translation introduces depth-dependent parallax and occlusion changes, whereas rotation and intrinsics remap rays at a fixed optical center. We therefore train a video model to generate canonical views while preserving the requested translation (Sec. 3.1). We then apply the remaining rotation and target intrinsics through reprojection, producing a target-view video that may contain missing regions. A target-camera-conditioned model completes these regions and refines the result (Sec. 3.2). Finally, we perform test-time scaling to further improve camera-control accuracy (Sec. 3.3).

3.1 Learning Video Generation in Canonical Views

To reduce the camera variation that the video model must learn, we train it to generate *canonical videos*, which are videos under simplified camera conditions. These videos provide the scene content for subsequent reprojection to the requested views. To make this reprojection independent of scene depth, we retain the target camera centers $\mathbf{o}_t$, thereby preserving the requested translation, and modify only orientation and intrinsics. Specifically, for a canonical camera $\mathcal{C}_t^{\text{can}} = (R_t^{\text{can}}, \mathbf{o}_t, \kappa^{\text{can}})$, a target pixel with homogeneous coordinates $\widetilde{\mathbf{p}} = (u, v, 1)^\top$ maps to the canonical view as

$$\widetilde{\mathbf{q}}_t \propto \kappa^{\text{can}}\big((R_t^{\text{can}})^\top R_t\, \kappa_t^{-1}(\widetilde{\mathbf{p}})\big)\,. \tag{1}$$

Here, κ_t^{-1} unprojects the target pixel to a camera ray, and $\widetilde{\mathbf{q}}_t$ denotes the corresponding canonical pixel in homogeneous coordinates. This mapping allows rotation and intrinsic changes to be applied wherever the canonical view covers the required rays.

The choice of canonical orientation R_t^{can} and projection κ^{can} determines which camera variations the generator must learn. We consider two canonical representations: level perspective views and equirectangular panoramas. Perspective views more closely match the conventional video domain used in pretraining, whereas panoramic generation requires adaptation to ERP geometry with more limited training data (Xia et al., 2025). At comparable pixel budgets, perspective views also concentrate resolution within a limited field of view, while panoramas provide full angular coverage. We examine their camera-control and visual-quality trade-offs in Sec. 4.2.

3.1.1 Level perspective views

We remove pitch and roll while retaining yaw, keeping the canonical camera aligned with the target's horizontal heading. Writing $R_t = R_y(\psi_t)R_x(\theta_t)R_z(\phi_t)$, we set $R_t^{\text{can}} = R_y(\psi_t)$ and use a fixed pinhole projection κ^{can}, with a horizontal FoV of 100° in our primary implementation. Reprojection then applies the target pitch, roll, and intrinsics.

We render canonical training videos from the PanFlow panoramas underlying PanShot (Zhang et al., 2026b), pairing each video with its corresponding canonical camera conditions. Our primary implementation initializes from Wan2.1 with UCPE camera conditioning (Wan et al., 2025; Zhang et al., 2026a), and we fine-tune both the video backbone and camera adapters on these paired data. UCPE conditions the DiT on relative ray frames and absolute latitude/up maps, while text enters through cross-attention. We denote this perspective variant as **CanoCam-P**.

3.1.2 Panoramic views

To provide source content across all viewing directions, we also generate equirectangular panoramas (ERP). Full-sphere coverage allows target rotations to be applied without exposing rays outside the source field of view, reducing the need for missing-region completion. We use the ERP projection for κ^{can} and, at inference, set $R_t^{\text{can}} = I$ while retaining the requested translation. All target rotation is therefore handled through reprojection.

For this panoramic variant, **CanoCam-E**, we initialize the generator from PanoWan (Xia et al., 2025) to reuse its learned panoramic prior. We then fine-tune the video backbone and UCPE camera adapters using ERP rays to incorporate camera translation control. Training uses native panoramic clips with their recovered camera poses, including their small rotations.

3.2 Reprojection and Target-Conditioned Refinement

3.2.1 Reprojection to Target Views

Using the mapping in Eq. 1, we compute the source pixel coordinates $\mathbf{q}_t(\mathbf{p})$ corresponding to each target pixel $\mathbf{p}$. We then sample the generated canonical video V^{can} at valid source locations using bicubic interpolation:

$$\widetilde{V}_t(\mathbf{p}) = V_t^{\text{can}}(\mathbf{q}_t(\mathbf{p}))\,. \tag{2}$$

We also construct a binary mask $H_t(\mathbf{p})$, which equals one for pixels without a valid source observation and zero otherwise. The same construction supports pinhole, unified-camera-model (UCM) (Mei & Rives, 2007), and ERP cameras through their respective projection mappings.

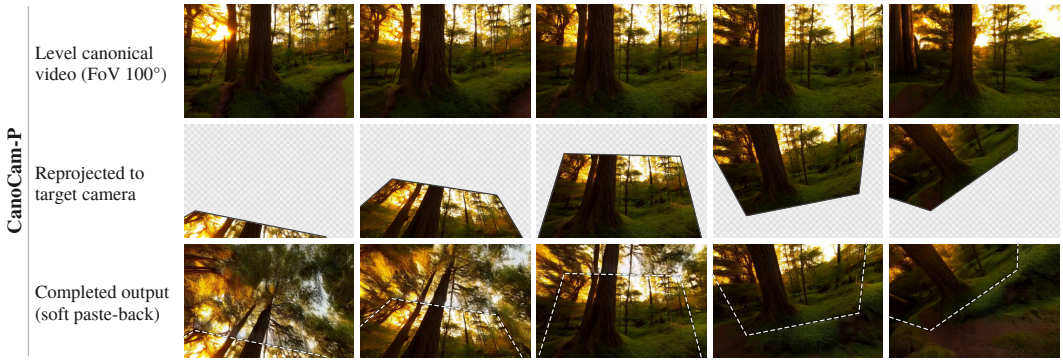

Figure 3: **Stages of CanoCam-P.** Rows show canonical views, reprojected observations, and refined outputs after blending. Checkerboard marks missing observations; dashed lines mark hole boundaries. Columns advance in time. No candidate selection is applied.

For level perspective videos, large rotations or wide target FoVs can expose regions outside the canonical field of view, leaving holes in the reprojected video. ERP videos provide full angular coverage, but their reprojected views remain affected by finite resolution and imperfections in the generated content. We therefore use a target-camera-conditioned video model for both representations to complete missing regions and refine the reprojected video.

3.2.2 Camera-Conditioned Completion and Refinement

The reprojected video provides scene content, but may not sufficiently constrain camera geometry when large regions are missing. We therefore use a camera-conditioned refinement model F_ϕ to produce a completed video V^{pred} consistent with the target camera poses and intrinsics:

$$V^{\text{pred}} = F_\phi(\widetilde{V}, H, c, \mathcal{C}; \eta). \tag{3}$$

Here, $\widetilde{V} = \{\widetilde{V}_t\}_{t=1}^{T}$ and $H = \{H_t\}_{t=1}^{T}$ collect the projected frames and masks; c and $\mathcal{C}$ are the text and target cameras, ϕ denotes model parameters, and η is sampling noise.

We concatenate the VAE-encoded observation and the hole mask at latent resolution with the noisy video latent along the channel dimension. We zero-initialize the added input projection weights and fine-tune the video backbone and camera adapters. Canonical generation and refinement use separate model weights, and both receive the text prompt.

For training, we reproject real videos into level canonical views and back, tracking validity to obtain partial observations and hole masks. The model reconstructs the original videos from these inputs. We vary the canonical FoV and include fully masked and fully observed examples. Flow matching supervises the full latent grid, enabling both completion and refinement (Appendix A).

3.2.3 Blending with Reprojected Observations

For perspective videos, we blend the predicted video with the reprojected observation to limit changes to observed content and smooth the transition around completed regions:

$$V^{\text{out}} = A \odot V^{\text{pred}} + (1 - A) \odot \widetilde{V}, \qquad A = \max(H, G_\sigma(\text{dilate}_r(H))) \,. \tag{4}$$

Here, r controls mask dilation and G_σ denotes Gaussian smoothing with standard deviation σ. The elementwise maximum with H makes holes use the prediction, while the soft mask A blends it with observed content near their boundaries; $\odot$ denotes elementwise multiplication. For ERP sources, we use the refined prediction directly as the final output, $V^{\text{out}} = V^{\text{pred}}$.

Figure 3 illustrates these stages over time.

3.3 Test-Time Scaling through Camera-Based Selection

Reprojection can propagate source orientation errors, and refinement may introduce further deviations. Their variation across samples motivates generating K complete candidates $\{V^{\text{out},(k)}\}_{k=1}^{K}$

with different random seeds under the same text and cameras. Following best-of-N sampling (Ma et al., 2025; Liu et al., 2025), we select the candidate whose estimated pitch and roll best match the request.

For each candidate, Perspective Fields (PF) (Jin et al., 2023) estimates pitch and roll in every output frame. Let (θ_t, ϕ_t) denote the requested pitch and roll from R_t, and let $(\hat{\theta}_t^{(k)}, \hat{\phi}_t^{(k)})$ denote the corresponding estimates for candidate k. We select

$$k^\star = \arg\min_{k\in\{1,\dots,K\}} \frac{1}{T}\sum_{t=1}^{T}\left[\left|\hat{\theta}_t^{(k)} - \theta_t\right| + \left|\mathrm{wrap}\left(\hat{\phi}_t^{(k)} - \phi_t\right)\right|\right], \tag{5}$$

where wrap maps angular differences to $[-180^\circ, 180^\circ)$. The selected output is the complete video $V^{\mathrm{out},(k^\star)}$. Selection requires the requested camera angles but no ground-truth video.

Total generation and scoring compute scales approximately linearly with K. The selection score measures estimated pitch and roll. We additionally use an independent orientation estimator, relative-pose metrics, and real-render controls to assess improvements beyond the selection score.

4 Experiments

4.1 Experimental Setup

Models and data. Our primary models build on Wan2.1-T2V-1.3B with UCPE camera adapters (Wan et al., 2025; Zhang et al., 2026a). We fine-tune the DiT and camera adapters, keeping the text encoder and VAE frozen. The perspective generator and refiner use 12,006 PanShot/PanFlow-derived clips (Zhang et al., 2026b); the panoramic generator adapts PanoWan (Xia et al., 2025) on 2,266 PanFlow ERP clips. CanoCam-P uses a completion-oriented refiner, while CanoCam-E uses a variant also trained on degraded observations. CanoCam-$\mathrm{E}_{\mathrm{proj}}$ denotes the panoramic projection before refinement. We also initialize both perspective stages from GimbalDiffusion's Wan2.2-5B checkpoint. We evaluate baselines using released checkpoints. Appendices A and D provide training and inference settings.

Benchmarks and metrics. SpatialVID-Extreme contains 138 requests with 49-frame outputs (Fortier-Chouinard et al., 2026). We evaluate pitch, roll, and up-vector errors with Perspective Fields (PF) (Jin et al., 2023), relative pose errors with VGGT (Wang et al., 2025b), and appearance with CLIP, FVD, FID, and Q-Align (Wu et al., 2024). GeoCalib (GC) (Veicht et al., 2024) provides orientation estimates independent of PF-based selection. Since GC is less reliable at extreme pitches, GC_{60} uses frames with requested $|\mathrm{pitch}| \le 60^\circ$, averaging within each video and then across the 131 eligible videos; it denotes pitch error unless specified otherwise. Prompt Entanglement evaluates first frames for 20 scenes at 19 static pitches (380 requests). PanShot tests pinhole and fisheye trajectories with UCPE's 81-frame protocol (Zhang et al., 2026a), using the 272 of 340 clips retained by its jitter filter. We report differences between GeoCalib estimates on generated and reference videos for vertical FoV, k_1, k_2, pitch, and roll; ViPE (Huang et al., 2025) trajectory errors; and the same appearance metrics. Additional pose evaluations use ViPE on SpatialVID-Extreme and VGGT on PanShot, covering CanoCam-$\mathrm{E}_{\mathrm{proj}}$ only for the panoramic variant. SpatialVID-Extreme uses eight seeds for UCPE, GimbalDiffusion 5B, and the primary CanoCam variants, and four for GimbalDiffusion UCPE. Other comparisons use one seed unless stated otherwise. Appendix B details prompts, protocols, data overlap, and checkpoint selection.

4.2 Camera Control across Viewpoints and Lenses

Control under extreme orientations. Table 1 compares CanoCam with UCPE and GimbalDiffusion on SpatialVID-Extreme. CanoCam-P reduces UCPE's PF pitch error from 16.09° to 9.18° and GC_{60} pitch error from 14.89° to 8.47°. It also achieves lower GC_{60} and relative rotation errors than GimbalDiffusion 5B, together with lower FVD and FID, while their PF pitch errors do not differ significantly. CanoCam-P obtains the lowest roll errors under both PF and GC_{60}. CanoCam-E achieves the lowest relative rotation error among the compared methods (10.59°), and its PF pitch error (9.29°) does not differ significantly from that of GimbalDiffusion 5B (8.58°). The benefit ex-

Table 1: **SpatialVID-Extreme without candidate selection.** N is the number of seed runs (4 or 8); $\pm$ gives the SD of their PF pitch means. Other entries average run-level scores; Q-Align (video quality) uses seeds 0–3 for every method. PF and GC_{60} errors are in degrees. Lower is better except for Q-Align; best and second-best means are **bold** and underlined, with ties sharing a rank.

		PF		GC_{60}		VGGT				
Method	N	Pitch	Roll	Pitch	Roll	Rot. (°)	Trans.	FVD	FID	Q-Align
UCPE	8	16.09±0.55	9.29	14.89	8.13	17.86	0.436	987	114.0	**0.733**
GimbalDiffusion UCPE	4	13.85±0.97	9.61	14.80	8.89	12.57	0.478	987	114.1	0.690
GimbalDiffusion 5B	8	**8.58**±0.73	7.56	11.52	6.57	14.50	0.732	1002	116.4	0.692
CanoCam-P (Ours)	8	9.18±0.76	**6.25**	**8.47**	**3.52**	13.27	**0.409**	**938**	**111.7**	0.731
CanoCam-E (Ours)	8	9.29±0.53	7.24	10.46	5.74	**10.59**	0.458	989	121.9	0.719

Table 2: **PanShot with pinhole and fisheye targets.** Following UCPE's protocol, every column uses the 272 of 340 test clips retained by its jitter filter. Lens and orientation errors compare GeoCalib estimates (radial model) on each generated video with those on the ground-truth video; FoV is the vertical FoV error, and FoV, pitch, and roll are in degrees. ViPE rotation sums per-frame angles in radians and differs in scale from Table 1; VGGT was not run on refined CanoCam-E outputs. Q-Align reports video quality. Lower is better except for CLIP and Q-Align; best and second-best scores are **bold** and underlined.

	Lens and orientation (GeoCalib)					ViPE			VGGT	Video quality			
Method	FoV	k_1	k_2	Pitch	Roll	Rot.	Trans.	CamMC	Rot.	FVD	FID	CLIP	Q-Align
UCPE	9.47	0.162	0.117	4.83	4.25	4.23	14.86	17.31	4.90	**454**	62.7	**26.41**	**0.913**
GimbalDiffusion UCPE	20.65	0.246	0.120	6.02	4.85	7.36	20.48	24.75	6.01	645	69.7	26.15	0.852
CanoCam-P (Ours)	**8.34**	**0.135**	**0.106**	**4.16**	3.91	3.24	13.00	14.76	**3.69**	477	**62.0**	26.19	0.875
CanoCam-E (Ours)	8.50	0.152	0.124	4.48	**3.90**	**3.18**	**12.28**	**14.12**	–	483	63.7	25.57	0.793

tends to the 5B backbone: on the four seeds shared by both 5B models, adapting GimbalDiffusion 5B to CanoCam reduces GC_{60} error from $11.38°$ to $7.91°$ (Appendix D).

Control across camera models. Table 2 evaluates pinhole and fisheye control on PanShot. Both CanoCam variants reduce ViPE rotation, translation, and CamMC errors relative to UCPE and GimbalDiffusion UCPE. CanoCam-E reduces UCPE's rotation error from 4.23 to 3.18 and CamMC from 17.31 to 14.12. VGGT also favors CanoCam-P over UCPE (3.69 versus 4.90 rotation error). Under UCPE's GeoCalib protocol, CanoCam-P obtains the lowest FoV, k_1, k_2, and pitch errors, and CanoCam-E the lowest roll error. The improvements over UCPE in FoV and pitch are small relative to GeoCalib's reference-footage errors (Appendix B). UCPE achieves the best FVD, CLIP, and Q-Align scores, and CanoCam-P the best FID.

Figure 4 shows qualitative comparisons on both benchmarks.

4.3 Ablation Studies and Analysis

Learning canonical views. Table 3 evaluates alternatives to CanoCam-P, using the full pipeline as the common reference. In *w/o canonical fine-tuning*, we use the pretrained UCPE generator at the same level canonical cameras. PF pitch error increases from $8.87°$ to $11.37°$, showing the benefit of adapting generation to canonical views. In *w/ target pitch*, we train and sample the source generator with the requested pitch retained and roll removed, increasing the error to $16.91°$. Both variants retain the same downstream refiner. These comparisons support restricting the camera variation assigned to source generation.

Canonical generation versus post-hoc correction. The *w/ post-hoc rectification* variant generates target views directly with UCPE, corrects their estimated orientation using GeoCalib, and applies the same refiner and blending operation. It obtains $13.67°$ PF pitch error and $25.14°$ relative rotation error, compared with $8.87°$ and $13.07°$ for CanoCam-P. Canonical generation therefore provides more effective source observations than this generate–rectify–complete alternative.

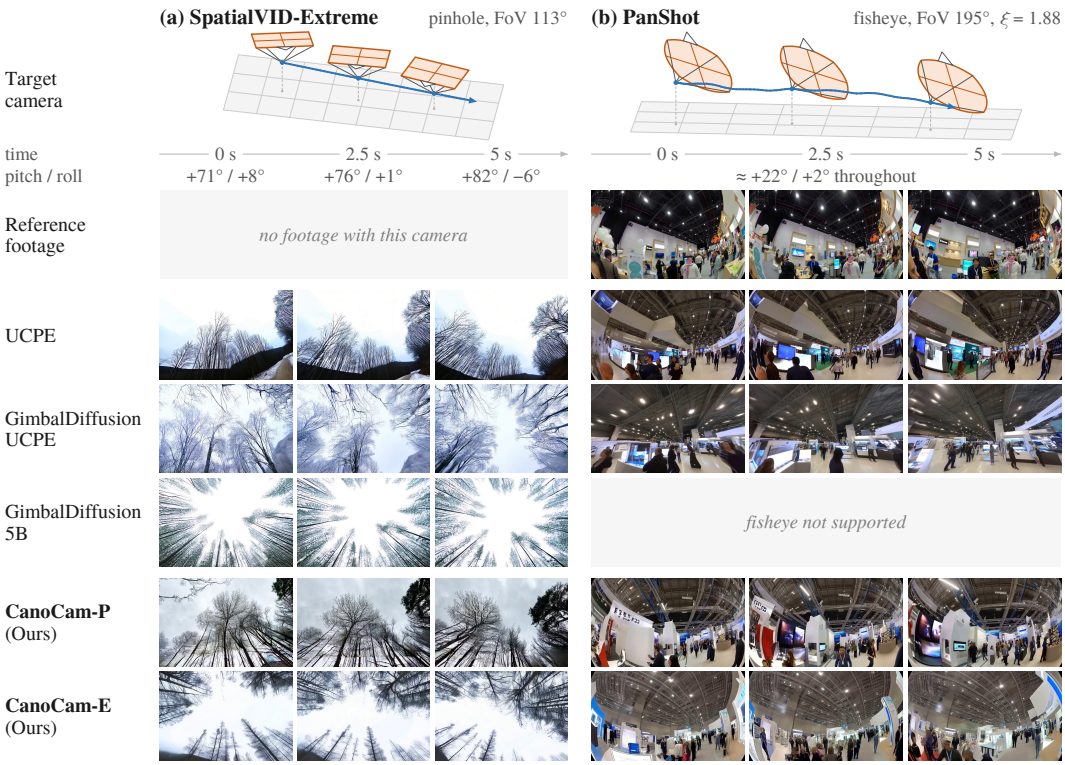

Figure 4: **Qualitative comparison.** Frames at 0, 2.5, and 5 s for (a) a SpatialVID-Extreme pinhole request (horizontal FoV 113°) that looks steeply upward, from 71° to 82° pitch, and (b) a PanShot fisheye request (horizontal FoV 195°, $\xi = 1.88$). The top row shows the target camera path with the requested pitch and roll, and the PanShot reference footage illustrates the requested camera. GimbalDiffusion 5B does not support fisheye cameras. All outputs are single samples without candidate selection; Tables 1 and 2 report average performance.

Table 3: **Ablation study of CanoCam-P on SpatialVID-Extreme.** All rows use seed 0. Source replacements keep the refiner fixed. The camera variant changes only the refiner's input condition. VACE uses the same reprojected observations and final blending. GeoCalib also guides rectification (†). Lower is better; best and second-best scores are **bold** and underlined, with ties sharing a rank.

Variant	PF pitch (°)	GC_{60} (°)	Rot. (°)	FVD
w/o canonical fine-tuning	11.37	10.82	18.48	<u>919</u>
w/ target pitch	16.91	15.20	16.21	958
w/ post-hoc rectification	13.67	13.23†	25.14	1023
w/o reprojected observations	15.21	15.17	17.05	1000
w/ canonical camera condition	15.68	11.10	18.41	959
w/ VACE refinement	13.46	11.29	17.36	**886**
w/o blending	<u>9.32</u>	<u>9.31</u>	<u>13.19</u>	941
CanoCam-P (full, Ours)	**8.87**	**8.30**	**13.07**	<u>919</u>

Camera-conditioned refinement. The *w/o reprojected observations* variant replaces the refiner's observation with a fully masked input while retaining its text and target camera conditions. PF pitch error rises to 15.21°, demonstrating the contribution of the canonical observation. In *w/ canonical camera condition*, we supply the canonical camera in place of the target camera to the same refiner, raising PF pitch error to 15.68°. The *w/ VACE refinement* variant uses VACE (Jiang et al., 2025) with the same reprojected observations and final blending. Its higher orientation errors, despite a lower FVD, show that a general completion model does not recover the same camera accuracy. Finally,

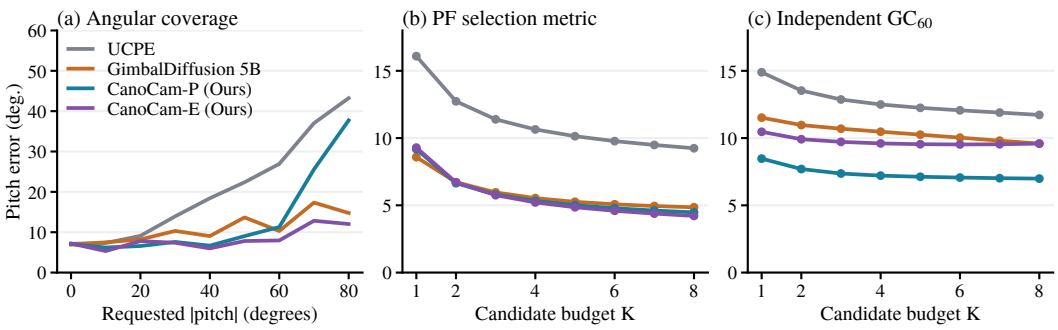

Figure 5: **Angular coverage and candidate budget.** (a) PF pitch error on Prompt Entanglement, averaging scenes and both pitch signs; the $\pm 90^\circ$ requests are omitted from this panel. (b,c) PF-based selection on SpatialVID-Extreme, averaged over all size-K subsets of eight candidates for every method; GC is not used for selection, and overlapping subsets are not independent trials.

w/o blending returns the refiner prediction directly. Its GC_{60} error increases from 8.30° to 9.31° and FVD from 919 to 941, supporting the retention of observed content.

Refining panoramic projections. Compared with CanoCam-E_{proj}, refinement improves CanoCam-E's Q-Align score on all three benchmarks, including an increase from 0.634 to 0.719 on SpatialVID-Extreme. This gain comes with higher PF and relative-pose errors and worse FVD/FID on SpatialVID-Extreme, while GC_{60} shows no significant change. On PanShot, refinement improves FVD and resolves four ViPE focal-length failures. Its lower full-set rotation error does not persist on the subset where ViPE succeeds for every method (Appendix C.1).

Perspective and panoramic generation. Figure 5a compares the two configurations on Prompt Entanglement through 80° absolute pitch. At 30°–50°, CanoCam-P achieves lower PF pitch error than GimbalDiffusion 5B (7.76° versus 11.01°). At 60°–80°, its error rises to 24.86°, while CanoCam-E achieves 10.94° versus 14.13° for GimbalDiffusion 5B, although this difference is not significant under scene-clustered resampling. Meanwhile, CanoCam-P obtains lower FVD and FID than CanoCam-E in Table 1. These results illustrate the trade-off between extreme-view control and visual quality across the two configurations.

4.4 Test-Time Scaling

Figure 5(b,c) evaluates selection across eight candidates per method. At $K = 8$, CanoCam-P and CanoCam-E achieve PF pitch errors of 4.48° and 4.21°, respectively, compared with 4.85° for GimbalDiffusion 5B; their differences from GimbalDiffusion 5B are not significant. Independent GC_{60} errors decrease from 8.47° to 6.99° for CanoCam-P and from 10.46° to 9.59° for CanoCam-E. The latter saturates after $K = 4$, despite further PF improvement. Selection therefore improves estimated orientation accuracy with approximately linear growth in generation and scoring compute. Appendix C.4 reports all budgets and computational costs.

5 Conclusion

CanoCam restricts the camera distribution assigned to generation: canonical videos retain translation and scene dynamics, reprojection applies the remaining camera transformations, and target-conditioned refinement completes and refines the result. Controlled comparisons support canonical generation and target-camera conditioning over the tested alternatives, including post-hoc correction, with control gains also observed on a second backbone. Perspective and panoramic sources offer different coverage and appearance trade-offs, while candidate selection further reduces estimated orientation error.

AI USE STATEMENT

We used large language models (LLMs) to improve the manuscript's clarity and readability and reduce its length to fit the 9-page limit. AI assistance was also used to automate the setup and execution of selected baseline methods following their official GitHub instructions, assist with visualizations, and submit jobs to GPU servers. The authors inspected the reproduced results and checked the evaluation metrics against those reported in the original papers. The core method was designed and implemented by the authors. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of the final manuscript.

REFERENCES

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22875–22889, 2025a.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. VD3D: Taming large video diffusion transformers for 3D camera control. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://snap-research.github.io/vd3d/.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14834–14844, 2025a.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. SynCamMaster: Synchronizing multi-camera video generation from diverse viewpoints. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. VideoPainter: Any-length video inpainting and editing with plug-and-play context control. In *SIGGRAPH 2025 Conference Papers*, pp. 1–12. Association for Computing Machinery, 2025. doi: 10.1145/3721238.3730673.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Ding Ding, Yueming Pan, Ruoyu Feng, Qi Dai, Kai Qiu, Jianmin Bao, Chong Luo, and Zhenzhong Chen. HomoGen: Enhanced video inpainting via homography propagation and diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22953–22962, 2025.

Frédéric Fortier-Chouinard, Yannick Hold-Geoffroy, Valentin Deschaintre, Matheus Gadelha, and Jean-François Lalonde. GimbalDiffusion: Gravity-aware camera control for video generation. In *European Conference on Computer Vision (ECCV)*, 2026. arXiv:2512.09112.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. CameraCtrl II: Dynamic scene exploration via camera-controlled video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13416–13426, 2025b.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022b.

Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3397–3406, 2022.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. CogVideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=rB6TpjAuSRy.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixé, and Sanja Fidler. ViPE: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025. URL https://research.nvidia.com/labs/toronto-ai/vipe/.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F. Fouhey. Perspective fields for single image camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17307–17316, 2023.

Min-Jung Kim, Jeongho Kim, Hoiyeong Jin, Junha Hyung, and Jaegul Choo. Infinite-homography as robust conditioning for camera-controlled video generation. *arXiv preprint arXiv:2512.17040*, 2025.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a. arXiv:2507.10496.

Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, Lin Cheng, and Xi Li. RealCam-I2V: Real-world image-to-video generation with interactive complex camera control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28785–28796, 2025b.

Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, et al. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8207–8216, 2024.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-t1: Test-time scaling for video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 18671–18681, 2025.

Rundong Luo, Matthew Wallingford, Ali Farhadi, Noah Snavely, and Wei-Chiu Ma. Beyond the frame: Generating 360° panoramic videos from perspective videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14336–14345, 2025.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Scaling inference time compute for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2523–2534, 2025. doi: 10.1109/CVPR52734.2025.00241.

Christopher Mei and Patrick Rives. Single view point omnidirectional camera calibration from planar grids. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3945–3950, 2007. doi: 10.1109/ROBOT.2007.364084.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Jing Tan, Shuai Yang, Tong Wu, Jingwen He, Yuwei Guo, Ziwei Liu, and Dahua Lin. Imagine360: Immersive 360 video generation from perspective anchor. In *Advances in Neural Information Processing Systems*, volume 38, pp. 139856–139874, 2025. doi: 10.52202/085713-4671.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/3526_ECCV_2020_paper.php.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. GeoCalib: Learning single-image calibration with geometric optimization. In *European Conference on Computer Vision (ECCV)*, 2024. doi: 10.1007/978-3-031-73661-2_1. arXiv:2409.06704.

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.

Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. PhysCtrl: Generative physics for controllable and physics-grounded video generation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167907–167932, 2025a. doi: 10.52202/085713-5598.

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 (CVPR)*, 2025b. arXiv:2503.11651.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 360DVD: Controllable panorama video generation with 360-degree video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6913–6923, 2024a.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b. doi: 10.1145/3641519.3657518.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *International Conference on Machine Learning*, pp. 54015–54029, 2024. URL https://proceedings.mlr.press/v235/wu24ah.html.

Yifei Xia, Shuchen Weng, Siqi Yang, Jingqi Liu, Chengxuan Zhu, Minggui Teng, Zijian Jia, Han Jiang, and Boxin Shi. PanoWan: Lifting diffusion video generation models to 360° with latitude/longitude-aware mechanisms. In *Advances in Neural Information Processing Systems*, volume 38, pp. 164289–164308, 2025. doi: 10.52202/085713-4949.

Xiaoda Yang, Jiayang Xu, Kaixuan Luan, Xinyu Zhan, Hongshun Qiu, Shijun Shi, Hao Li, Shuai Yang, Li Zhang, Checheng Yu, et al. Omnicam: Unified multimodal video generation via camera control. *arXiv preprint arXiv:2504.02312*, 2025a.

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 *The Thirteenth International Conference on Learning Representations*, 2025b.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3613256.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Cruz Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38027–38037, 2026a.

Cheng Zhang, Hanwen Liang, Donny Y. Chen, Qianyi Wu, Konstantinos N. Plataniotis, Camilo Cruz Gambardella, and Jianfei Cai. PanFlow: Decoupled motion control for panoramic video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12385–12393, 2026b. doi: 10.1609/aaai.v40i15.38231. URL https://ojs.aaai.org/index.php/AAAI/article/view/38231.

A Implementation Details

Canonical generators and data. The perspective generator retains translation and yaw, removes pitch and roll, and uses a 100° pinhole horizontal FoV. Rotations map camera coordinates to a gravity-referenced world frame with $+y$ pointing down; translations are divided by the source clip's near-depth estimate. We render level views from PanFlow panoramas at the corresponding PanShot frame indices. The 12,006 training clips comprise 12,000 non-validation clips ranked by maximum absolute pitch or roll within 3°–80°, plus six near-180° cases. The refiner uses the same clips with their original target cameras. The panoramic generator uses 2,266 parent clips, initializes from PanoWan merged into Wan2.1, and is fine-tuned with UCPE camera conditioning and ERP rays. Training retains the recovered poses; inference sets orientation to identity.

Refinement inputs. The input projection receives 16 noisy latent channels, four packed mask channels, and 16 observation-latent channels. Added weights are zero-initialized. Missing pixels are set to zero in the normalized $[-1, 1]$ image range before VAE encoding. Training observations use target-to-canonical-to-target reprojection with canonical FoVs of 90°, 100°, or 110°. Approximately 10% of examples are fully masked and 5% fully observed; other draws prefer hole fractions below 0.9. Invalid target rays outside the UCM image circle are also masked. The evaluated perspective refiner fills these pixels, and PanShot metrics use the unmasked output.

Refinement-oriented variant for panoramic projections. Projected ERP sources have no holes, but their detail is limited by source resolution and generation quality. For CanoCam-E, we therefore train a refinement-oriented variant with an additional degraded-observation task. Reprojected videos are downsampled, blurred, and compressed, with mask value 0.5 indicating observed but degraded pixels. To simulate a projected panoramic source, we resize the real panorama of a training clip to the 576×1152 generation resolution with 8-bit quantization and apply Gaussian blur with standard deviation drawn uniformly from $[0, 0.5]$ ERP pixels. We then project it into the clip's target camera, apply a small color jitter with probability 0.5, and re-encode it with H.264 at a constant rate factor drawn from 19 to 29. We precompute two degraded observations for each of 2,965 pinhole training clips. Each training example follows the completion setup above with probability 0.5. With probability 0.35, it provides the complete degraded observation with all mask channels set to 0.5. The remaining examples combine the degraded observation with a hole pattern, using mask value 1 in holes and 0.5 elsewhere. The target is always the original video. At inference, all projected pixels receive mask value 0.5, and CanoCam-E uses the prediction without blending.

CanoCam-P uses the completion-oriented variant. With CanoCam-P sources and the same blending, the refinement-oriented variant does not improve seed-0 SpatialVID-Extreme accuracy. PF pitch error changes from 8.87° to 9.93° ($+1.06^\circ$, 95% interval $[-0.12, 2.30]$), and relative rotation error from 13.07° to 13.73° ($+0.66^\circ$, $[-0.25, 1.91]$).

Optimization. For latent x_0, noise ϵ, and noise level $\sigma(t)$, all components use the inherited flow-matching objective:

$$x_\sigma = (1 - \sigma)x_0 + \sigma\epsilon, \qquad \mathcal{L} = \mathbb{E}\left[w(t)\|v_\theta(x_\sigma, t, \mathcal{S}) - (\epsilon - x_0)\|_2^2\right], \tag{6}$$

where $\mathcal{S}$ contains text, cameras, and any observation and mask. The loss covers the full latent grid. The scheduler uses 1,000 timesteps, shift 5, uniform sampling over the shifted grid, and its inherited weight $w(t)$. AdamW uses learning rate 10^{-5} for the DiT and camera adapters, 10^{-4} for the widened input, weight decay 0.01, 200 warm-up steps (50 for the refinement-oriented variant), and gradient clipping at 1.0. Training uses bf16 and ZeRO stage 1 with fp32 master weights; the VAE and text encoder are frozen. Table 4 gives component-specific settings.

Inference and projection. SpatialVID-Extreme outputs have 49 frames at 480×720; PanShot outputs have 81 frames at 480×832. We use 50 sampling steps and guidance scale 5. ERP sources use 576×1152 on SpatialVID-Extreme and Prompt Entanglement, and 448×896 on PanShot. Candidate k uses the benchmark seed plus $1000k$ for both the source generator and the refiner of either variant. CanoCam-E refinement samples from noise with 50 steps at the output resolution, conditioned on the full projected view. Perspective blending applies four binary-dilation iterations and Gaussian smoothing with $\sigma = 4$ pixels, followed by the maximum with the original hole mask. ERP projection uses zero padding at the seam, affecting approximately 0.030% of target pixels on

Table 4: Training settings for the 1.3B models. All use 81 frames and one clip per GPU, giving effective batch size 8, or 16 for the CanoCam-E variant. Resolution is height $\times$ width. Cosine decay spans steps 1,000–2,000 for the generators, 2,000–4,000 for the CanoCam-P variant of the refiner, and 600–1,200 for the CanoCam-E variant, reaching 0.1 times the initial learning rate. The CanoCam-E variant uses 50 warm-up steps and mixes completion examples with degraded observations of 2,965 pinhole clips.

Component	Clips	Resolution	GPUs $\times$ accum.	Steps
Perspective generator	12,006	480×832	8×1	2,000
Panoramic generator	2,266	448×896	8×1	2,000
Refiner (CanoCam-P variant)	12,006	480×832	4×2	2,500
Refiner (CanoCam-E variant)	12,006	480×832	8×2	1,200

average. A ray–projection round-trip check reveals a one-pixel offset per ERP axis; the evaluated path projects directly from ERP and uses the same ray convention in training and inference.

B Evaluation Protocol and Data Exposure

Benchmarks and prompts. SpatialVID-Extreme uses the same 138 requests in every run, with target horizontal FoV approximately $112.62°$. PF, VGGT, FVD/FID (Unterthiner et al., 2018; Heusel et al., 2017), and CLIP (Radford et al., 2021) follow the benchmark implementations. Multi-seed FVD/FID entries average separately computed run-level scores. GimbalDiffusion 5B outputs are resized from width 736 to 720. GimbalDiffusion 5B uses eight seeds on SpatialVID-Extreme; comparisons with CanoCam-5B use their four shared seeds. Baselines are local reproductions of released checkpoints. On this benchmark, UCPE and GimbalDiffusion use the benchmark's look-down negative prompt, while our primary 1.3B variants use UCPE's default. Changing UCPE to the default prompt gives seed-0 PF error $16.27°$, versus $15.84°$ with the benchmark prompt. PanShot uses the UCPE default throughout. CanoCam-E results include refinement. CanoCam-$\mathrm{E}_{\mathrm{proj}}$ denotes the projected panoramic source without refinement; we report it where it isolates the refinement stage or where an estimate was not computed for refined outputs.

Prompt Entanglement evaluates the first frame for 20 scenes at 19 pitches from $-90°$ to $90°$. Figure 5a omits the $\pm 90°$ cases (340 requests); Table 7 retains all 380. PanShot follows UCPE's evaluation and uses, for every metric, the 272 of 340 clips retained by its jitter filter. ViPE uses the original stride-1 protocol. Rotation error sums per-frame angles in radians over 81 frames; translation and CamMC use the official sequence aggregation. FVD uses no center crop. For lens and orientation errors, GeoCalib with the radial camera model and intrinsics shared across frames is run on every fourth frame of each generated and ground-truth video; absolute differences in vertical FoV, k_1, k_2, pitch, and roll are averaged within each video and then across videos. Our UCPE regeneration reproduces the pose errors reported by UCPE (4.23, 14.86, and 17.31 versus 4.12, 15.21, and 17.59), whereas its GeoCalib lens and orientation errors are higher than reported (e.g., FoV error $9.47°$ versus $8.22°$), so we compare all methods under this common protocol.

Interpreting PanShot calibration errors. Under the common GeoCalib protocol, GimbalDiffusion UCPE has a mean estimated vertical FoV of $77.4°$, versus $61.5°$ for the reference footage. CanoCam-P reduces UCPE's FoV and pitch errors by $1.1°$ and $0.7°$, respectively, with paired 95% intervals excluding zero. These margins are small relative to GeoCalib's error on reference footage, where it deviates from the known pitch by about $7°$ on average.

Data overlap and model selection. The 128 perspective validation clips share all 126 parent panoramas with training. Panoramic training and validation have disjoint parent clips, but share 90 of 116 validation source videos. These development splits therefore do not establish source-level generalization. Checkpoints, including that of the refinement-oriented variant, canonical FoV, and ERP inference resolution were informed by SpatialVID-Extreme seed 0; additional seeds measure sampling variation conditional on these choices.

PanShot has no identical training/test clip in the available audit, but 228 test clips share source videos with UCPE training. Removing this overlap leaves 112 clips, or 92 after filtering. Additional

PanoWan/GimbalDiffusion metadata identify six overlapping test sources (36 clips, including 16 with temporal overlap). Excluding all known source overlap leaves 100 clips, or 81 after filtering. We call this the *strict known-source-disjoint* subset; incomplete pretraining metadata prevents a claim of complete independence. No shared IDs were found for SpatialVID-Extreme or Prompt Entanglement, although a source-video audit of the former is unavailable.

Uncertainty. We compute paired 95% bootstrap intervals by resampling source videos for PanShot (10,000 draws, retaining all associated clips), scenes for Prompt Entanglement (10,000 draws), and requests for SpatialVID-Extreme (5,000 draws over per-request seed means). Intervals are not corrected for multiple comparisons. Seed standard deviations describe variation in run means; single-seed intervals do not measure training-run variability.

C Additional Evaluation

C.1 Refining Panoramic Projections

Table 5: **Effect of refining panoramic projections.** CanoCam-E$_{\mathrm{proj}}$ and CanoCam-E use the same panoramic sources and seeds; CanoCam-E additionally applies the refiner. Differences are CanoCam-E minus CanoCam-E$_{\mathrm{proj}}$ with paired 95% intervals (Appendix B). SpatialVID-Extreme uses eight seeds, including Q-Align; FVD and FID are set-level scores, for which we give the number of seeds on which CanoCam-E is worse. Prompt Entanglement CLIP uses the benchmark's pitch-adjusted prompt. PanShot pose errors are sequence sums with rotation in radians; focal failures count clips on which ViPE returns a negative focal length, and the ViPE-valid subset keeps the 262 clips on which ViPE succeeds for UCPE, GimbalDiffusion UCPE, CanoCam-P, CanoCam-E, and CanoCam-E$_{\mathrm{proj}}$. PanShot FVD, FID, CLIP, and Q-Align use the 272 clips of Table 2. Lower is better except for CLIP and Q-Align. The better mean in each pair is **bold**; this does not indicate statistical significance.

Metric	CanoCam-E$_{\mathrm{proj}}$ (Ours)	CanoCam-E (Ours)	Difference [95% interval]
SpatialVID-Extreme			
PF pitch (°)	**7.94**	9.29	$+1.35$ $[0.69, 2.11]$
GC_{60} (°)	10.57	**10.46**	-0.11 $[-0.96, 0.66]$
VGGT Rot. (°)	**8.06**	10.59	$+2.53$ $[2.01, 3.09]$
VGGT Trans.	**0.391**	0.458	$+0.066$ $[0.049, 0.084]$
CLIP	19.73	**19.74**	$+0.00$ $[-0.19, 0.19]$
Q-Align	0.634	**0.719**	$+0.084$ $[0.075, 0.094]$
FVD	**958**	989	$+30$ (7 of 8 seeds)
FID	**119.0**	121.9	$+2.9$ (8 of 8 seeds)
Prompt Entanglement			
PF pitch, all (°)	**10.82**	11.39	$+0.57$ $[-0.84, 2.10]$
PF pitch, 60°–90° (°)	**16.18**	17.57	$+1.39$ $[-1.87, 4.81]$
CLIP	**22.29**	21.43	-0.85 $[-1.18, -0.53]$
Q-Align	0.666	**0.800**	$+0.134$ $[0.121, 0.148]$
PanShot			
ViPE focal failures (of 272)	4	**0**	–
ViPE Rot., 272 clips	3.38	**3.18**	-0.20 $[-1.33, 0.69]$
ViPE Rot., 262 ViPE-valid clips	**2.71**	3.24	$+0.53$ $[0.09, 1.04]$
ViPE Trans., 272 clips	12.62	**12.28**	-0.34 $[-1.83, 1.13]$
CamMC, 272 clips	14.59	**14.12**	-0.46 $[-2.57, 1.37]$
CLIP	25.27	**25.57**	$+0.31$ $[0.23, 0.38]$
Q-Align	0.558	**0.793**	$+0.234$ $[0.217, 0.253]$
FVD	519	**483**	-36
FID	65.9	**63.7**	-2.2

Table 5 isolates the refiner in CanoCam-E by comparing it with CanoCam-E$_{\mathrm{proj}}$, which uses the same panoramic sources and seeds. Refinement increases Q-Align by 0.084 on SpatialVID-Extreme, 0.134 on Prompt Entanglement, and 0.234 on PanShot, and adds fine texture to projected views. The

refined output retains the projected view's global layout and horizon, but can add content in weakly constrained regions, such as solid structures in a featureless sky. This content can change between frames. On SpatialVID-Extreme, refinement increases PF pitch, relative rotation, and translation errors. FID and FVD increase for most seeds, while GC_{60} and CLIP show no significant change. On Prompt Entanglement, the change in PF pitch error is not significant, and CLIP decreases. On PanShot, refinement improves FVD and FID and removes the four ViPE focal-length failures of CanoCam-E_{proj}. Mean rotation error decreases over all 272 clips but increases slightly on the 262 clips where ViPE succeeds for every method.

Among the ranked methods in Table 1, CanoCam-E has the lowest SpatialVID-Extreme relative rotation error. Its PF pitch error does not differ significantly from that of GimbalDiffusion 5B with eight seeds (paired difference $+0.71°$, 95% interval $[-0.64, 2.02]$), whereas its GC_{60}, rotation, and translation errors are lower by $1.05°$ ($[0.01, 2.16]$), $3.91°$ ($[2.77, 5.14]$), and 0.275 ($[0.238, 0.310]$), respectively. On Prompt Entanglement, CanoCam-E has a lower mean pitch error than every baseline in Table 7 ($11.39°$ versus $13.89°$ for GimbalDiffusion 5B), but the difference from GimbalDiffusion 5B ($-2.50°$, $[-5.33, 0.41]$) is not significant under scene-clustered resampling. On PanShot, CanoCam-E reduces UCPE's rotation error from 4.23 to 3.18 (-1.05, $[-1.56, -0.54]$); on the 262 ViPE-valid clips, the reduction is 0.79 ($[0.28, 1.31]$).

C.2 Appearance, Angular Coverage, and Data Exposure

Table 6: Additional SpatialVID-Extreme metrics. Up-vector error and CLIP average eight seeds, except GimbalDiffusion UCPE and CanoCam-5B (four each); optical flow uses seed 0. Up-vector error is lower-is-better; CLIP is higher-is-better. Flow magnitude measures motion and has no preferred direction. CanoCam-E_{proj} is the panoramic projection without refinement; – marks optical flow that was not computed. Best and second-best scores among complete methods are **bold** and underlined; the projection-only row and flow magnitudes are not ranked.

Method	Up error (°)	CLIP	Flow (px/frame)
UCPE	18.67	**23.07**	10.19
GimbalDiffusion UCPE	16.74	<u>21.72</u>	–
GimbalDiffusion 5B	10.99	20.75	11.22
CanoCam-P (Ours)	<u>10.82</u>	20.45	12.25
CanoCam-E (Ours)	11.41	19.74	–
CanoCam-5B (Ours)	**9.68**	19.75	–
CanoCam-E_{proj} (Ours)	9.75	19.73	14.15

Table 7: PF pitch error (°) on Prompt Entanglement, using one seed per method. The overall mean weights all 19 signed pitches equally, including $\pm 90°$. Columns group requests by absolute pitch. CanoCam-E_{proj} is the panoramic projection without refinement. PanoWan uses the same 576×1152 resolution and target projection as CanoCam-E_{proj}, also without refinement. Lower is better; best and second-best scores among complete methods (upper block) are **bold** and underlined.

Method	Overall	0°–20°	30°–50°	60°–90°
UCPE	25.10	8.00	18.24	40.92
GimbalDiffusion 5B	<u>13.89</u>	7.67	11.01	<u>19.95</u>
CanoCam-P (Ours)	17.46	**6.53**	<u>7.76</u>	31.56
CanoCam-E (Ours)	**11.39**	<u>6.69</u>	**7.07**	**17.57**
CanoCam-E_{proj} (Ours)	10.82	6.65	7.16	16.18
PanoWan	13.94	10.07	10.48	18.95

On SpatialVID-Extreme (seeds 0–3, Table 1), CanoCam-P and UCPE have similar Q-Align scores (paired difference -0.003, 95% interval $[-0.019, 0.013]$). Refinement raises CanoCam-E's Q-Align from 0.635 for CanoCam-E_{proj} to 0.719, close to CanoCam-P (paired difference -0.012, $[-0.024, 0.001]$). On PanShot and Prompt Entanglement, CanoCam-E remains below CanoCam-P in Q-Align (0.793 versus 0.875, and 0.800 versus 0.824). CanoCam-E retains higher FVD and FID than CanoCam-P (Table 1). Table 6 reports up-vector error, CLIP, and motion magnitude. Q-

Align (Wu et al., 2024) uses 4-fps sampling; motion magnitude uses consecutive-frame RAFT-large flow (Teed & Deng, 2020) at 8 fps and measures motion amount, not realism. Compared with PanoWan at the same 576×1152 resolution, CanoCam-E_{proj} reduces SpatialVID-Extreme PF pitch error from 13.53° to 7.94° and translation error from 0.721 to 0.391; the prior uses one seed and CanoCam-E_{proj} eight. PanoWan retains lower FVD (898 versus 958). On Prompt Entanglement, the pitch reduction is 3.11° (95% interval $[0.67, 5.62]$). Both are evaluated without refinement; the comparison changes both fine-tuning and camera conditioning.

Table 8: PanShot rotation error (sequence sum, radians) under the original ViPE protocol. Source-disjoint subsets exclude videos used in UCPE training or all identified training sources (the strict subset), respectively. "Src." counts YouTube source videos. The lens subsets are defined by UCM distortion ξ; source and lens subsets overlap. CanoCam-E_{proj} is the panoramic projection without refinement and is not ranked. Lower is better; best and second-best scores among complete methods are **bold** and underlined within each subset.

Subset	Clips	Src.	UCPE	CanoCam-P (Ours)	CanoCam-E (Ours)	CanoCam-E_{proj} (Ours)
All evaluated clips	272	45	4.23	<u>3.24</u>	**3.18**	3.38
Disjoint from UCPE training	92	16	5.10	**2.87**	<u>3.68</u>	3.31
Strict known-source-disjoint	81	13	5.50	**3.03**	<u>3.90</u>	3.57
Pinhole ($\xi = 0$)	69	36	<u>4.93</u>	5.10	**2.90**	5.52
Fisheye ($0 < \xi \leq 1$)	56	32	4.30	**2.29**	<u>2.70</u>	2.18
Fisheye ($\xi > 1$)	147	44	3.87	**2.73**	<u>3.50</u>	2.84

Table 8 retains the same PanShot filtering protocol within each subset. CanoCam-P, CanoCam-E, and CanoCam-E_{proj} all reduce mean rotation error relative to UCPE on fisheye cameras and on the strict known-source-disjoint subset. On pinhole cameras, CanoCam-P and CanoCam-E_{proj} have higher means than UCPE, whereas CanoCam-E has a lower mean (2.90 versus 4.93); three of the four ViPE focal failures of CanoCam-E_{proj} (Table 5) occur on pinhole clips. On the strict subset, the reductions relative to UCPE are 2.47 for CanoCam-P (95% interval $[1.35, 4.05]$), 1.60 for CanoCam-E ($[0.75, 2.70]$), and 1.93 for CanoCam-E_{proj} ($[0.73, 3.31]$), based on 13 source videos.

Table 9: PanShot pinhole subset ($\xi = 0$). The 69 pinhole clips among the 272 clips of Table 2 allow comparison with GimbalDiffusion 5B, which we evaluate only on pinhole requests. ViPE errors are sequence sums with rotation in radians. FVD and FID are computed on these 69 clips and are comparable only within this table. Q-Align reports video quality. CanoCam-E_{proj} is the panoramic projection without refinement and is not ranked. Lower is better except for CLIP and Q-Align; best and second-best scores are **bold** and underlined.

	ViPE						
Method	Rot.	Trans.	CamMC	FVD	FID	CLIP	Q-Align
UCPE	<u>4.93</u>	<u>13.45</u>	<u>16.30</u>	**1048**	**107.8**	**26.47**	**0.889**
GimbalDiffusion UCPE	5.31	18.50	21.17	1173	111.8	26.02	0.813
GimbalDiffusion 5B	12.54	45.63	52.27	1242	<u>108.3</u>	25.73	0.682
CanoCam-P (Ours)	5.10	**12.02**	**15.23**	<u>1066</u>	109.7	<u>26.06</u>	<u>0.837</u>
CanoCam-E (Ours)	**2.90**	15.08	16.69	1147	111.7	25.21	0.691
CanoCam-E_{proj} (Ours)	5.52	14.37	17.85	1096	121.4	24.67	0.441

On the pinhole clips (Table 9), GimbalDiffusion 5B has the highest ViPE rotation, translation, and CamMC errors and the highest FVD among the ranked methods, while its FID is second only to UCPE. CanoCam-P obtains higher CLIP and Q-Align scores than GimbalDiffusion 5B (Q-Align difference $+0.154$, 95% interval $[0.120, 0.190]$), and CanoCam-E has the lowest rotation error.

C.3 Sensitivity to Camera Estimation

Table 10 compares VGGT and ViPE using identical generated videos. ViPE's default initialization uses GeoCalib; nominal initialization uses $f = (W/2)/\tan(\mathrm{FoV}/2)$, with focal length still optimized. No additional clips are excluded according to estimation outcomes. CanoCam-E_{proj} has the

Table 10: Trajectory errors across evaluators. SpatialVID-Extreme uses 138 requests, seed 0 for UCPE and CanoCam, and released GimbalDiffusion videos; rotation is in degrees and translation is normalized. PanShot uses the same 272 clips as Table 2; both errors are sequence sums, with rotation in radians. Original ViPE uses GeoCalib focal initialization; nominal ViPE uses the requested rectified focal length. Reference footage has known cameras; its SpatialVID-Extreme translations do not match the requests and are omitted. CanoCam-E_{proj} is the panoramic projection without refinement; – marks estimates that were not computed for refined CanoCam-E. Lower is better; best and second-best scores among complete methods are **bold** and underlined within each dataset. Reference and projection-only rows are not ranked.

	VGGT		ViPE (original)		ViPE (nominal)	
Method	Rot.	Trans.	Rot.	Trans.	Rot.	Trans.
SpatialVID-Extreme						
Reference footage	6.15	–	2.15	–	2.15	–
UCPE	18.36	0.423	17.53	0.346	18.00	0.360
GimbalDiffusion UCPE	11.81	0.460	**8.92**	0.350	**8.79**	0.317
GimbalDiffusion 5B	15.47	0.747	13.30	0.695	11.95	0.707
CanoCam-P (Ours)	13.07	**0.395**	11.69	**0.297**	11.06	**0.286**
CanoCam-E (Ours)	**10.40**	0.428	–	–	–	–
CanoCam-E_{proj} (Ours)	8.94	0.390	5.29	0.271	4.37	0.222
PanShot						
Reference footage	3.21	16.52	1.29	6.17	1.26	6.35
UCPE	4.90	**20.52**	4.23	14.86	4.05	14.73
GimbalDiffusion UCPE	6.01	23.55	7.36	20.48	6.46	20.46
CanoCam-P (Ours)	**3.69**	21.19	3.24	13.00	**2.44**	**12.04**
CanoCam-E (Ours)	–	–	**3.18**	**12.28**	–	–
CanoCam-E_{proj} (Ours)	3.14	15.69	3.38	12.62	2.35	11.43

lowest SpatialVID-Extreme rotation error among generated methods under all three estimates. ViPE was not run on refined CanoCam-E outputs; its seed-0 VGGT rotation error ($10.40°$) remains below those of GimbalDiffusion UCPE ($11.81°$) and CanoCam-P ($13.07°$). On PanShot, CanoCam-P and CanoCam-E_{proj} improve rotation over UCPE under VGGT and both ViPE initializations, and CanoCam-E does so under original ViPE, while CanoCam-P's translation advantage is not consistent across estimators. Reference-footage errors characterize estimator sensitivity and are not exact lower bounds. The released GimbalDiffusion UCPE videos used here are its best-rotation draw, distinct from the four-seed mean in Table 1.

Table 11: First-frame pitch error (°) on Prompt Entanglement, using one seed per method. The reference uses real panoramas rendered at the requested pitch. Parenthesized GC values are unreliable at extreme pitches, as shown by the reference error, and are not used to rank methods. CanoCam-E_{proj} is the panoramic projection without refinement. Lower is better; best and second-best scores among complete methods are **bold** and underlined. Reference and projection-only rows are not ranked.

	$\lvert\mathrm{pitch}\rvert \le 50°$		$\lvert\mathrm{pitch}\rvert \ge 60°$	
Method	PF	GC	PF	GC
Real-render reference	6.2	4.8	12.3	(43.0)
UCPE	13.6	13.1	40.9	(51.8)
GimbalDiffusion UCPE	15.0	14.8	32.3	(52.3)
GimbalDiffusion 5B	9.5	8.7	19.9	(50.2)
CanoCam-P (Ours)	7.2	**5.6**	31.6	(52.1)
CanoCam-E (Ours)	**6.9**	7.4	**17.6**	(48.6)
CanoCam-E_{proj} (Ours)	6.9	7.5	16.2	(56.8)

GC_{60} averages frames with requested $\lvert\mathrm{pitch}\rvert \le 60°$ within each video, then averages the 131 eligible videos. In the real-render control of Table 11, GC's reference pitch error reaches $43.0°$ near vertical views. We therefore use PF to assess extreme-pitch control. At $\lvert\mathrm{pitch}\rvert \le 50°$, both estimators favor CanoCam over UCPE. GC favors CanoCam-P over CanoCam-E, whereas PF does not resolve their difference.

C.4 Candidate Selection and Computational Cost

Table 12: **Selection using PF pitch-plus-roll error on SpatialVID-Extreme.** Each entry averages all size-K subsets of the candidate pool, then averages requests. "Subsets" counts these overlapping combinations. The selector chooses a whole video; GC is not used for selection. GimbalDiffusion UCPE has four candidates; – marks GC scores that were not computed. CanoCam-E_{proj} is the panoramic projection without refinement. Errors are in degrees; lower is better. Best and second-best scores among complete methods at each K are **bold** and underlined, with ties sharing a rank. The projection-only rows are not ranked.

Method	K	Subsets	PF pitch	GC_{60}	GC, all	VGGT Rot.
UCPE	1	8	16.09	14.89	18.49	17.86
UCPE	4	70	10.64	12.50	15.78	17.55
UCPE	8	1	9.24	11.72	14.74	17.39
GimbalDiffusion UCPE	1	4	13.85	14.80	–	12.57
GimbalDiffusion UCPE	4	1	8.17	–	–	12.30
GimbalDiffusion 5B	1	8	**8.58**	11.52	17.09	14.50
GimbalDiffusion 5B	4	70	5.54	10.47	15.80	13.74
GimbalDiffusion 5B	8	1	4.85	9.59	14.83	13.38
CanoCam-P (Ours)	1	8	9.18	**8.47**	**13.33**	13.27
CanoCam-P (Ours)	4	70	5.32	**7.21**	**11.85**	12.81
CanoCam-P (Ours)	8	1	4.48	**6.99**	**11.83**	12.84
CanoCam-E (Ours)	1	8	9.29	10.46	15.78	**10.59**
CanoCam-E (Ours)	4	70	**5.21**	9.60	14.61	**10.41**
CanoCam-E (Ours)	8	1	**4.21**	9.59	14.49	**10.30**
CanoCam-E_{proj} (Ours)	1	8	7.94	10.57	16.88	8.06
CanoCam-E_{proj} (Ours)	4	70	4.38	9.30	15.76	8.01
CanoCam-E_{proj} (Ours)	8	1	3.61	9.47	15.52	8.17

At each budget K, Figure 5 and Table 12 average all size-K subsets of the available candidates. Every method in Figure 5(b,c) has eight candidates, so $K = 4$ averages 70 overlapping subsets; these are not independent repetitions. Selection improves PF and independent GC scores, but its effect on relative pose is smaller and less consistent. For CanoCam-E, GC_{60} error decreases from 10.46° at $K = 1$ to 9.60° at $K = 4$ (paired gain 0.86°, 95% interval $[0.38, 1.40]$), with little further change at $K = 8$. Its VGGT rotation error is 10.59° at $K = 1$ and 10.30° at $K = 8$ (gain 0.29°, $[-0.62, 1.08]$). Refinement increases PF error at every budget. At $K = 4$, CanoCam-E exceeds CanoCam-E_{proj} by 0.83° ($[0.40, 1.46]$), and its difference from GimbalDiffusion 5B (-0.32°, $[-1.25, 0.79]$) is not significant.

Selecting among four real-panorama reprojections also reduces PF error from 4.96° to 3.01° and GC_{60} from 6.45° to 5.58°. This illustrates content-dependent estimator error. Because the panorama pool was screened using level-view PF/GC estimates and can retain source tilt, its gain cannot be subtracted from generated-video gains to estimate physical camera improvement.

Generating one CanoCam-P candidate takes approximately 63.2 B200-seconds, comprising 30.42 for generation, 1.55 for reprojection, and 31.24 for refinement, plus 1.8 CPU-seconds for blending. CanoCam-E takes approximately 115 B200-seconds, including 83.8 for generation and projection and about 31 for refinement. PF scoring adds about 0.44 GPU-seconds per candidate. GimbalDiffusion 5B generation takes 12.3 B200-seconds. Timing excludes model loading and combines stage measurements from different drivers, so these costs do not support an exact matched-compute comparison. Fine-tuning costs are approximately 44.6 B200-hours for CanoCam-P and 67 for CanoCam-E (19.6 for the panoramic generator and about 47.6 for its refinement-oriented variant, estimated from checkpoint times), excluding preprocessing and inherited pretraining.

D Extension to a 5B Backbone

Implementation. Both stages initialize from GimbalDiffusion's Wan2.2-5B model with its Plücker camera adapter and are fully fine-tuned on the same 12,006 clips and 128 validation IDs as CanoCam-P. The refinement input has 48 noisy latent, four mask, and 48 observation-latent chan-

nels; added weights are zero-initialized. Training uses 41 frames sampled every other frame, effective batch size 8, and resolutions 480×736 for the source and 480×832 for refinement. The source trains for 3,000 steps and refinement for 2,500. Learning rates match Appendix A, with 100 warm-up steps and cosine decay to one tenth of the initial rate, starting at steps 1,500 and 2,000, respectively.

Inference uses 49 frames, 50 steps, and guidance scale 5. Both pretrained and fine-tuned 5B sources are centrally cropped from width 736 to 720, with canonical horizontal FoV approximately $112.62°$. Translation is normalized by maximum trajectory displacement with a floor of 0.1, following GimbalDiffusion, rather than UCPE's near-depth convention.

Table 13: Source and refiner size on SpatialVID-Extreme, seed 0. Pretrained 5B uses GimbalDiffusion 5B at canonical cameras without fine-tuning; canonical sources are fine-tuned. Both 5B sources use identical cropping and FoV; the 1.3B source uses $100°$. Angular errors are in degrees, translation error is normalized, and GC_{60} uses requested $|\mathrm{pitch}| \leq 60°$.

Source	Refiner	PF pitch	GC_{60}	Rot.	Trans.
Pretrained 5B	1.3B	10.05	10.68	14.26	0.709
Pretrained 5B	5B	8.62	9.21	16.38	0.722
Canonical 5B	1.3B	8.75	9.46	11.61	0.608
Canonical 5B	5B	7.86	7.48	13.34	0.622
Canonical 1.3B	1.3B	8.87	8.30	13.07	0.395
Canonical 1.3B	5B	8.66	7.69	14.73	0.402

Source and refinement ablations. With the 5B refiner fixed, canonical fine-tuning reduces GC_{60} error from $9.21°$ to $7.48°$, rotation error from $16.38°$ to $13.34°$, and translation error from 0.722 to 0.622 (Table 13). The benefit of a larger refiner is mixed. With the canonical 5B source, GC improves while rotation error increases from $11.61°$ to $13.34°$. Over four seeds on SpatialVID-Extreme, CanoCam-5B obtains PF pitch and roll errors of $7.99° \pm 0.52°$ and $5.97°$, GC_{60} pitch and roll errors of $7.91°$ and $3.63°$, relative rotation and translation errors of $13.77°$ and 0.645, FVD 1047, FID 114.7, and Q-Align 0.698. On the same seeds, GimbalDiffusion 5B has a GC_{60} pitch error of $11.38°$. CanoCam-5B improves GC and translation over GimbalDiffusion 5B in every seed pair, but does not consistently outperform CanoCam-P. The comparison also changes initialization, conditioning, training, and FoV, so it does not isolate model size.

E Qualitative Comparisons and Failure Cases

All examples use candidate $k = 0$. Representative SpatialVID-Extreme cases are nearest the median five-method mean PF error within each maximum-absolute-pitch bin; the five methods are UCPE, GimbalDiffusion 5B, the refiner without observations, CanoCam-P, and CanoCam-E_{proj}. Ties use sample ID. Figures 3 and 4 instead show requests chosen so that the processing stages and the differences between methods are clearly visible. Figure 3 uses a SpatialVID-Extreme pinhole request with horizontal FoV $113°$ and shows frames 0, 12, 24, 36, and 48. The last two rows of Figure 6 instead maximize CanoCam-P and CanoCam-E_{proj} PF error, respectively.

Both variants can propagate source errors; limited perspective coverage and panoramic angular resolution remain distinct failure modes. Refinement and blending may leave inconsistent structures or visible seams, and refinement of panoramic projections can add content in weakly constrained regions. Viewpoint frequency, requested pitch, and missing coverage covary in our evaluation, so these experiments do not isolate rarity as a causal explanation for control error.

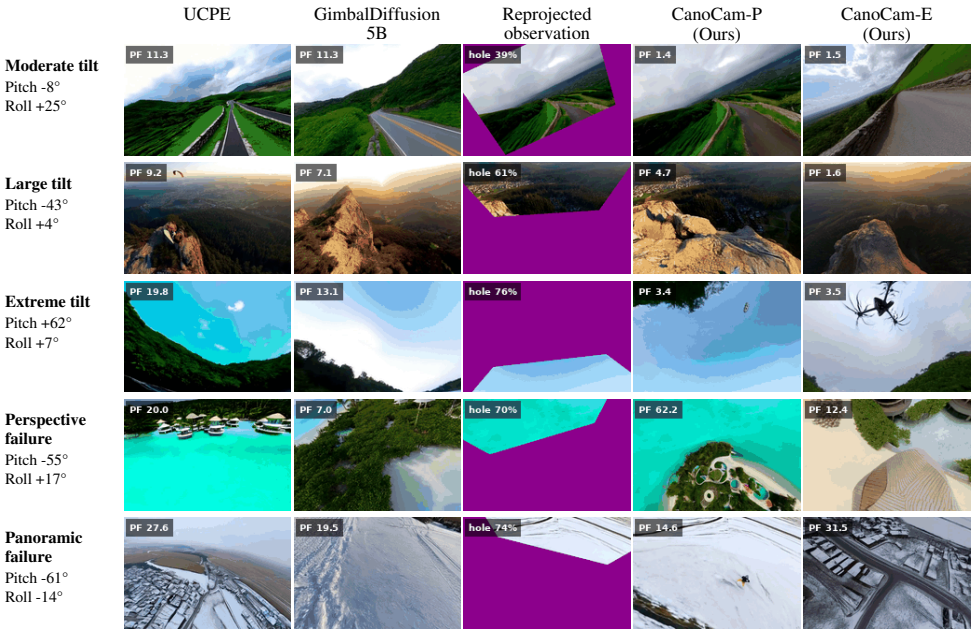

Figure 6: **Representative views and failure cases on SpatialVID-Extreme.** All columns share the prompt, camera request, and frame index (24). PF annotations report video-level pitch error; magenta marks missing observations. Extreme views can lose detail or introduce semantic artifacts. The CanoCam-E column shows CanoCam-E_{proj}, the projection before refinement. The perspective failure also exhibits estimator disagreement (GC_{60} error $7.75°$); the panoramic failure has large errors under both estimators.

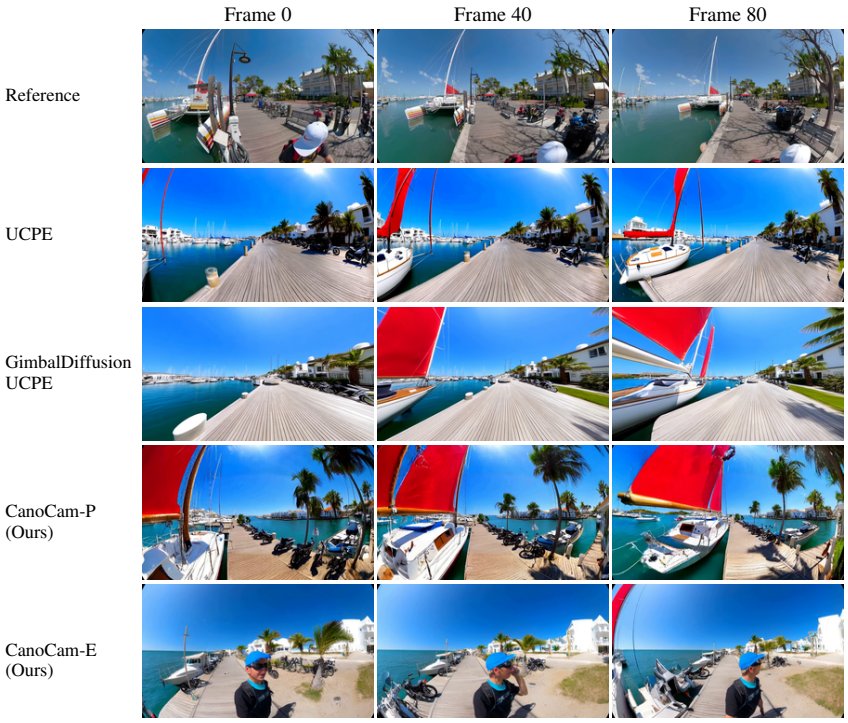

Figure 7: **A PanShot fisheye trajectory.** Frames 0, 40, and 80 use target horizontal FoV 200° and $\xi = 1.69$. This case is nearest the median CanoCam-P–UCPE rotation difference in the strict known-source-disjoint, $\xi > 1$ subset. The CanoCam-E row shows CanoCam-$\mathrm{E}_{\mathrm{proj}}$, the projection before refinement. Reference footage illustrates the camera request; text-conditioned generation need not reconstruct its scene.