CRePE: Curved Ray Expectation Positional Encoding for Unified-Camera-Controlled Video Generation

Seonghyun Jin*[1], Youngmin Kim*‡[1,2], Sunwoo Park*[1], and Jong Chul Ye†[1]

[1]Kim Jaechul Graduate School of Artificial Intelligence, KAIST, Seoul, Republic of Korea
[2]Korea University, Seoul, Republic of Korea
{jinotter3,sunwoo_p,jong.ye}@kaist.ac.kr; zeromin03@korea.ac.kr

Abstract. Video world models should predict future appearance in a way that remains consistent with 3D scene structure, camera motion, and lens geometry. Existing attention-level camera encodings, however, either describe each token only by its viewing ray—without locating scene content along that ray—or assume pinhole projection, limiting camera control under wide-angle and fisheye lenses. We introduce **Curved Ray Expectation Positional Encoding (CRePE)**, which represents each image token as a depth-aware distribution along its Unified Camera Model (UCM) ray and integrates the expected rotary positional phasor along the curved path this distribution traces when projected into each query view. CRePE is realized through a lightweight **Geometric Attention Adapter** on a frozen video diffusion transformer, with pseudo radial-distance supervision from a monocular geometry foundation model serving as a stabilizing anchor rather than an inference-time input. CRePE improves camera-control, lens, and orientation fidelity across pinhole, wide-angle, and fisheye settings, and transfers zero-shot to unseen real fisheye and diverse pinhole videos. Through **Radial MixForcing**, the same positional pathway further accepts externally supplied radial maps, enabling scene-geometry-conditioned generation and source-video motion transfer that follow the supplied geometry more faithfully than dedicated depth-conditioned baselines. CRePE thus offers a compact interface that unifies camera control, implicit 3D scene state, and external geometry control for video world models.

Keywords: video world models · camera control · geometric positional encoding · fisheye video generation

1 Introduction

World models for visual prediction increasingly use video diffusion transformers as scalable scene simulators [22, 32]. For such a model, camera control is not merely

⋆ Equal contribution. †Corresponding author. ‡Work performed during a KAIST internship.

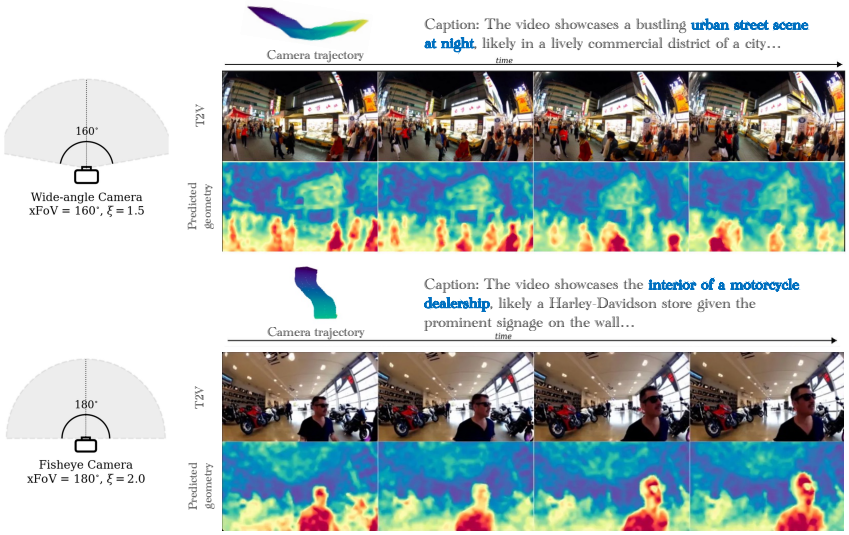

Fig. 1: Camera-controlled text-to-video generation under non-pinhole lenses. CRePE receives only a prompt, target trajectory, and lens parameters; no source video, reference image, or external geometry is used. Its internal token-wise radial distributions organize 3D scene position while the generated video follows wide-angle (top) and fisheye (bottom) camera motion.

an animation handle: it tests whether appearance changes remain compatible with an observer moving through a coherent 3D scene. Camera-conditioned video generation [3] therefore requires models to synthesize videos that follow specified camera trajectories and lens configurations. Recent studies have expanded the scope of camera conditioning from direct camera conditioning to attention-level positional encoding [28, 34]. By incorporating camera geometry directly into positional encoding, video diffusion transformers [15, 24] can more explicitly model the relative positional relationships among latent tokens under view changes. In particular, UCPE [34] encodes the viewing ray of each token based on a unified central-camera representation, demonstrating the possibility of camera control under the Unified Camera Model (UCM), including wide-angle and fisheye cameras. This marks an important step toward handling diverse lens models within a unified positional-encoding framework.

However, relative-ray-based positional encoding still has a limitation. A viewing ray indicates the direction associated with a token, but it does not specify where the underlying scene content lies along that ray. When the camera moves, token correspondences depend not only on ray direction but also on the scene position along the ray. Thus, while ray-only encoding provides a stable way to inject camera geometry, it remains limited in representing more refined positional relationships coupled with scene geometry. This limitation becomes more pronounced for wide-angle and fisheye cameras, where projection geometry changes nonlinearly.

To address this problem, we propose **Curved Ray Expectation Positional Encoding (CRePE)**. Instead of representing each token only by its viewing ray, CRePE models it as a depth-aware positional distribution along the source ray, reflecting possible scene positions associated with the token. CRePE is designed to be compatible with the Unified Camera Model, naturally capturing the nonlinear projection geometry that arises under diverse lens configurations. This allows the positional encoding to account for how tokens may align with the underlying scene structure and to refine token relationships under view changes.

To apply CRePE to camera-conditioned video generation, we introduce a **Geometric Attention Adapter** for frozen video DiTs. The adapter predicts token-wise scene distance and incorporates it into the positional relationships used in attention. Pseudo radial-distance supervision from a monocular geometry foundation model serves only as an auxiliary geometric anchor, helping the predicted geometry signal remain stable during long training. Our supervision-free controls show that the positional formulation itself supplies the initial camera-control gain; the anchor prevents the learned radial channel from drifting rather than injecting geometry at inference. Through this design, CRePE extends ray-only positional encoding into a scene-geometry-aware positional encoding while preserving the adapter-based camera-conditioning structure.

We evaluate CRePE against camera-conditioned video generation baselines [2, 34] and isolate the encoding through supervision-free controls. CRePE improves camera-control, lens, orientation, and perceptual metrics on PanShot, and transfers without tuning to real fisheye KITTI-360 and diverse RealCam-Vid videos. A Depth-Concat control shows that access to the same pseudo geometry is insufficient: radial distance must enter attention as a query-dependent phase. Inspired by YonoSplat [33], we further introduce **Radial MixForcing**, allowing the same positional pathway to accept external geometry for scene-geometry-conditioned generation and source-video motion transfer. Quantitative DAVIS and non-pinhole PanShot experiments verify adherence to the supplied maps.

Figure 1 illustrates representative wide-angle and fisheye camera-conditioned generations together with CRePE's internal radial-distance maps. Table 1 summarizes how CRePE differs from UCPE and RayRoPE-style endpoint PE along camera-model support, ray-position modeling, geometric grounding, and external-control capability.

Our contributions are summarized as follows:

- We propose **Curved Ray Expectation Positional Encoding**, a UCM-compatible encoding that integrates positional phase over a token's log-radial uncertainty after transport into each query camera.
- We apply CRePE through a **Geometric Attention Adapter** for frozen video DiTs, with pseudo radial supervision used as a long-training stabilizer rather than an inference-time input.
- Supervision-free K ablations, affine-path analysis, and pinhole/non-pinhole stratification separate non-affine integration from extra samples or pseudo-depth access; robustness tests cover alternate teachers, dynamic regions, and two zero-shot datasets.

– **Radial MixForcing** exposes the same pathway for external 3D control and yields competitive or better depth adherence with markedly lower temporal scale jitter than dedicated depth-conditioned video baselines.

2 Related Work

Camera-conditioned video generation. Camera-conditioned video models control generated motion using pose embeddings, camera trajectories, rendered geometric guidance, or dedicated camera branches [2, 5, 27, 31, 35]. ReCamMaster [2] is a strong external baseline for trajectory-conditioned generation, but its formulation is tied to a pinhole setting. UCPE [34] instead conditions video diffusion with unified-camera rays and is therefore the closest baseline for our setting. CRePE builds on the UCPE interface but adds a token-wise ray-position variable inside attention.

Geometric positional encoding. Rotary positional encoding [21] makes attention depend on relative position. Multi-view variants replace image-grid position with camera-aware quantities: CaPE [9], GTA [14], and PRoPE [11] encode relative camera geometry; raymap methods concatenate ray origins or directions to features [1,4,19,20]; RayRoPE [28] predicts a token-wise ray-position interval and applies expected RoPE after projection. Our RayRoPE-style endpoint PE ablation follows this endpoint-interval approximation. CRePE differs by using UCM projection for unified central cameras and by integrating the phasor over a curved projected path rather than a single endpoint-defined interval.

Video world models and geometric control. Recent video generators increasingly serve as predictive world models with explicit camera or 3D control [22, 32, 36]. Depth-conditioned systems inject dense structural features into a generative backbone [8, 13, 29, 30, 37]. CRePE instead treats 3D scene position as part of the relative positional relation used by attention. This distinction makes camera-only prediction and optional external geometry intervention share one compact interface.

Central-camera geometry and radial distance. Wide-angle and fisheye videos are better described by central-camera models such as UCM than by pinhole intrinsics. Radial distance—distance from the camera center along the viewing ray—is a natural scalar for such cameras because it remains defined regardless of lens distortion. Recent geometric foundation models provide dense monocular scene geometry across diverse cameras [10, 17, 25, 26]. We use these predictions only as pseudo radial-distance anchors for CRePE's latent ray-position head.

Table 1 compares CRePE with UCPE and a RayRoPE-style endpoint PE design across the main geometry-aware positional-encoding axes.

Table 1: Geometry-aware positional encodings for camera-conditioned video DiTs. We compare across five design axes: support for non-pinhole UCM projection, learnable per-token ray-position parameterization, integration over the curved projected path induced by ray-position uncertainty under non-pinhole projection, external grounding of the ray-position prediction via pseudo-GT supervision, and exposure of a joint camera+depth control interface at inference. "–" marks axes not applicable to a method without a learnable depth axis.

Property	UCPE [34]	RayRoPE-style PE [28]	CRePE (ours)
Non-pinhole / UCM projection	✓	✗	✓
Ray-position-adaptive phase	✗	✓	✓
Curved projected-path integration	–	✗	✓
Pseudo-GT geometric grounding	–	✗	✓
Joint camera+depth control interface	✗	✗	✓

3 Method

3.1 Overview

We start from a frozen video diffusion transformer with a UCPE-style camera adapter. For each token at layer ℓ, source frame s, and patch p, the adapter has feature $\boldsymbol{h}_{\ell,s,p}$. CRePE adds a geometry head g_ϕ that predicts a log radial-distance center μ and interval width σ:

$$(\mu_{\ell,s,p}, \sigma_{\ell,s,p}) = g_\phi(\boldsymbol{h}_{\ell,s,p}). \tag{1}$$

These predictions define a token-wise radial-distance distribution. CRePE converts this distribution into expected RoPE modulation coefficients by lifting radial-distance breakpoints along the source ray and projecting them into each query camera. The resulting coefficients modulate attention in a selected set of middle DiT layers. All other adapter layers keep UCPE's relative-ray modulation. Figure 2 gives an overview of this pipeline.

3.2 Curved-Ray Expected RoPE

CRePE interprets the predicted interval as a uniform distribution in log radial distance. Specifically, for each source token we define

$$z = \log r \sim \mathcal{U}(\mu - |\sigma|, \mu + |\sigma|),$$

where r denotes the radial distance measured from the source camera center along the token's UCM ray. We then discretize this log-radial interval with K uniformly spaced breakpoints and exponentiate them:

$$\begin{aligned} z_k &= \mu - |\sigma| + \frac{k-1}{K-1}(2|\sigma|), \\ r_k &= \exp(z_k), \qquad k = 1, \ldots, K. \end{aligned} \tag{2}$$

Each r_k is lifted along the source UCM ray, transformed into the query camera, and projected through the query UCM to obtain a query-side coordinate $\tilde{\mathbf{x}}^{(k)}$.

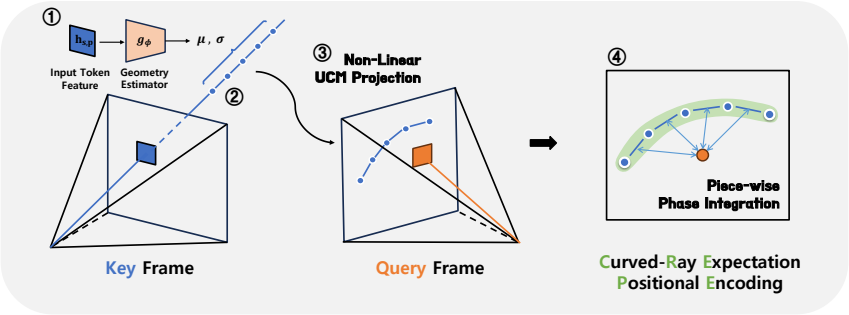

Fig. 2: CRePE. A shared geometry head g_ϕ predicts a log-distance center μ and uncertainty $|\sigma|$ from each key-side token feature $\mathbf{h}_{s,p}$, defining K breakpoints $\{r_k\}$ along the source viewing ray (①). The breakpoints are unprojected to 3D via the source UCM model $\Phi_{\xi_s}^{\mathrm{UCM}}$ (②), transported into the query frame and forward-projected via $\Pi_{\xi_q}^{\mathrm{UCM}}$, tracing a curved projected path $\{\tilde{\mathbf{x}}^{(k)}\}$ in the query view (③). CRePE aggregates the expected RoPE phasor along this trajectory by analytically integrating each of the $K-1$ piecewise-linear segments (④), yielding a token-level expected RoPE modulation $\bar{\rho}_{f,c}$ that captures ray-position uncertainty under non-pinhole projection.

For a fixed RoPE frequency ω_f and coordinate channel c, define $\theta_{f,c}^{(k)} = \omega_f \tilde{x}_c^{(k)}$. CRePE treats consecutive breakpoints as piecewise-linear phase segments and integrates the phasor analytically:

$$\begin{aligned} \bar{c}_{f,c} &= \frac{1}{K-1} \sum_{k=1}^{K-1} \frac{\sin \theta_{f,c}^{(k+1)} - \sin \theta_{f,c}^{(k)}}{\theta_{f,c}^{(k+1)} - \theta_{f,c}^{(k)}}, \\ \bar{s}_{f,c} &= \frac{1}{K-1} \sum_{k=1}^{K-1} \frac{\cos \theta_{f,c}^{(k)} - \cos \theta_{f,c}^{(k+1)}}{\theta_{f,c}^{(k+1)} - \theta_{f,c}^{(k)}}. \end{aligned} \tag{3}$$

The pair $(\bar{c}_{f,c}, \bar{s}_{f,c})$ forms the expected rotary-modulation coefficient for that coordinate and frequency. Because the expected phasor can have magnitude below one, this coefficient is an expected RoPE modulation rather than a strictly orthonormal rotation. Eq. (3) therefore approximates

$$\mathbb{E}_{z \sim \mathcal{U}(\mu - |\sigma|, \mu + |\sigma|)} \left[\exp(i \omega_f \tilde{x}_c(z)) \right],$$

where $\tilde{x}(z)$ is the query-view projection of the 3D point obtained by lifting radial distance $r = \exp(z)$ along the source UCM ray. Importantly, the expectation is taken over log radial distance, not over image coordinates or metric distance directly. CRePE thus computes the expected phasor along the curved projected path, rather than applying RoPE at a single averaged position $\exp(i\omega_f \mathbb{E}[\tilde{x}_c])$. When $K = 2$, the path reduces to an endpoint approximation; with $K > 2$, CRePE captures the curvature induced by UCM projection. We use $K = 5$ by default.

For numerical stability, the closed-form segment integral uses its continuous small-phase limit when the endpoint phase difference approaches zero.

CRePE inside attention. We apply the expected phasor to the key-side camera-conditioned token in the geometric attention branch. Let $i = (q, p_q)$ denote a query token in query frame q, and let $j = (s, p_s)$ denote a key/value token in source frame s. For a CRePE-enabled layer ℓ, the geometry head predicts $(\mu_{\ell,s,p_s}, \sigma_{\ell,s,p_s})$ from the key-side feature h_{ℓ,s,p_s}. Given the relative camera transform from source frame s to query frame q, Eq. (2)–(3) produces a query-conditioned expected rotary-modulation coefficient

$$\bar{m}_{\ell,q\leftarrow s,p_s} = \{(\bar{c}_{f,c}, \bar{s}_{f,c})\}_{f,c}.$$

We convert this coefficient into the corresponding block-diagonal modulation matrix $\bar{M}_{\ell,q\leftarrow s,p_s}$ and use it to modulate the key feature:

$$\begin{aligned} Q_{\ell,i} &= W_Q h_{\ell,i}, & V_{\ell,j} &= W_V h_{\ell,j}, \\ K^{\text{crepe}}_{\ell,j} &= \bar{M}_{\ell,q\leftarrow s,p_s} W_K h_{\ell,j}. \end{aligned}$$

The geometric attention score is then

$$\begin{aligned} A^{\text{crepe}}_{\ell,i,j} &= \text{softmax}_j \left(\frac{Q^{\top}_{\ell,i} K^{\text{crepe}}_{\ell,j}}{\sqrt{d}} \right), \\ O^{\text{crepe}}_{\ell,i} &= \sum_j A^{\text{crepe}}_{\ell,i,j} V_{\ell,j}. \end{aligned}$$

The output of this branch is passed through the same zero-initialized output projection used by the adapter and added residually to the frozen DiT feature. CRePE is used only in the selected middle layers; the remaining layers keep the UCPE relative-ray modulation. This preserves the stable ray-only camera signal while injecting depth-aware projected-path modulation where radial-distance information is most recoverable.

3.3 Geometric Attention Adapter and Radial-Distance Supervision

The ray-position head in Eq. (1) affects attention directly, so incorrect predictions can become a harmful positional shortcut. We therefore supervise it with pseudo radial-distance maps from a monocular geometry foundation model [17]. We first apply validity filtering to the raw metric pseudo distance $\hat{r}^{\text{m}}$: a target is invalid if UniK3D fails, is marked unreliable by our validity filters, returns a non-finite or non-positive value, falls outside the supported distance range, or predicts $\hat{r}^{\text{m}} > r_{\max} = 20\,\text{m}$. Invalid or far-field values are not clipped into pseudo ground truth. Only valid metric targets are normalized by a per-clip near-distance statistic, pooled to the token grid, and used as normalized targets $\hat{r}$ in the loss. From (μ, σ) we decode the predicted normalized radial distance $\bar{r} = \exp(\mu)$ and an uncertainty scale s derived from the log interval. The auxiliary loss is

$$\mathcal{L}_{\text{rad}} = \frac{1}{|\mathcal{T}|} \sum_{(c,p)\in\mathcal{T}} \left(\frac{|\bar{r}_{c,p} - \hat{r}_{c,p}|}{s_{c,p}} + \alpha \log s_{c,p} \right), \tag{4}$$

where $\mathcal{T}$ is the valid token set. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{rad}} \mathcal{L}_{\text{rad}}. \tag{5}$$

Here $\mathcal{L}_{\text{diff}}$ is the backbone diffusion objective and $\lambda_{\text{rad}} = 10^{-3}$. We supervise all CRePE-modulated layers and keep the video DiT backbone frozen. True UniK3D failures affect only 0.061% of pixels; after the intentional 20 m far-field cut, 56.3% of pixels (55.7% of tokens) remain valid targets. The remaining rays are always represented by the model prediction. Section 4.5 shows that this weak anchor primarily prevents drift under long optimization.

3.4 Radial MixForcing

Radial MixForcing is a training-time teacher-substitution algorithm for the CRePE ray-position pathway. Its purpose is to teach the adapter to accept externally supplied radial maps while preserving a safe camera-only fallback. Let $(\mu_{\text{pred}}, \sigma_{\text{pred}})$ be the radial head output, let $\hat{r}^{\text{m}}$ be the raw metric pseudo radial distance, let $\hat{r}$ be its per-clip-normalized value when valid, let $\mu_{\text{gt}} = \log \hat{r}$, and let M be a sampled teacher-substitution mask. We define pseudo-GT validity on the raw metric output as

$$\begin{aligned} V_{\text{gt}} &= \mathbf{1}[\text{valid}_{\text{UniK3D}}(\hat{r}^{\text{m}}) \wedge 0 < \hat{r}^{\text{m}} \leq r_{\max}], \\ r_{\max} &= 20\,\text{m}. \end{aligned} \tag{6}$$

Distances above the validity threshold are therefore treated as unknown, not as valid targets clipped to 20 m. CRePE consumes the effective radial distribution

$$(\mu_{\text{eff}}, \sigma_{\text{eff}}) = \begin{cases} (\mu_{\text{gt}}, \sigma_T) & \text{if } M \wedge V_{\text{gt}}, \\ (\mu_{\text{pred}}, \sigma_{\text{pred}}) & \text{otherwise}, \end{cases} \tag{7}$$

where σ_T is a narrow teacher interval. The second branch covers every invalid ray: if the estimate is unreliable, non-finite, non-positive, outside range, or above 20 m, no teacher geometry is injected. The auxiliary loss in Eq. (4) is likewise applied only to valid tokens. At inference, an external map follows the same filtering and normalization; invalid or missing rays fall back to $(\mu_{\text{pred}}, \sigma_{\text{pred}})$. We use the supervised CRePE model for camera-only evaluation and a MixForcing-trained model for external-map control. Training uses $\sigma_T = 0.1$; a 20× inference sweep over $[0.05, 1.0]$ changes CamMC by only 3.9%, with the best video quality in the stable 0.05–0.3 range.

3.5 Layer Placement

Applying CRePE to every layer is unnecessary and can degrade visual quality. We probe the frozen DiT layers for recoverable radial-distance information and find that middle layers provide the best tradeoff between geometric control and generation quality. Our default model applies CRePE to the middle 10 blocks

and retains UCPE modulation elsewhere. Table 3 reports the placement ablation over contiguous windows: the middle window attains the best average rank and the strongest pose, lens, and orientation performance, whereas early and late windows are competitive only on isolated video-quality metrics such as FVD and FID.

4 Experiments

4.1 Setup

Backbone and training. We use Wan2.1-T2V-1.3B [24] as the frozen video backbone and train only the camera adapter for 10K AdamW steps on 81-frame, 480×832 clips (global batch 8, learning rate 10^{-4}). CRePE is applied to the middle 10 blocks with $K = 5$ and radial supervision at the middle layer.

Dataset. We use the released PanShot train/test partition, with pinhole, wide-angle, and fisheye settings [34]. We evaluate all 340 test clips without filtering by camera-motion magnitude. Clips are sampled by temporal position; dense attention therefore observes offsets from adjacent pairs (median rotation 0.91°) through the full clip span (median 16.4°). Pseudo radial maps from UniK3D [17] are used only for training supervision or optional external-map interventions and are filtered before normalization.

Metrics. Following UCPE [34], we recover relative pose from generated videos with ViPE [6] and estimate lens and orientation with GeoCalib [23]. We report rotation error (RotErr), scale-aligned translation error (TransErr), CamMC, radial-distortion errors (k_1, k_2), pitch error, gravity error, and up-vector error. All methods use the same fixed estimators, so these scores are comparable proxies that may inherit estimator bias. We evaluate generation quality with FVD, FVD-C, FID, Inception Score, VBench metrics, and CLIP-based image/text scores.

4.2 Baselines

The main evaluation compares against **ReCamMaster** [2] and **UCPE** [34]. Because ReCamMaster is originally formulated for video re-rendering, we use it as an external camera-control baseline under the same recovered-camera evaluation protocol; UCPE remains the architecture-matched baseline with the same Wan backbone and adapter interface but without CRePE or radial-distance supervision. We evaluate **RayRoPE-style endpoint PE** [28] in a separate positional-encoding ablation, where it isolates the difference between endpoint-based expected RoPE and CRePE's UCM-compatible curved projected-path integration.

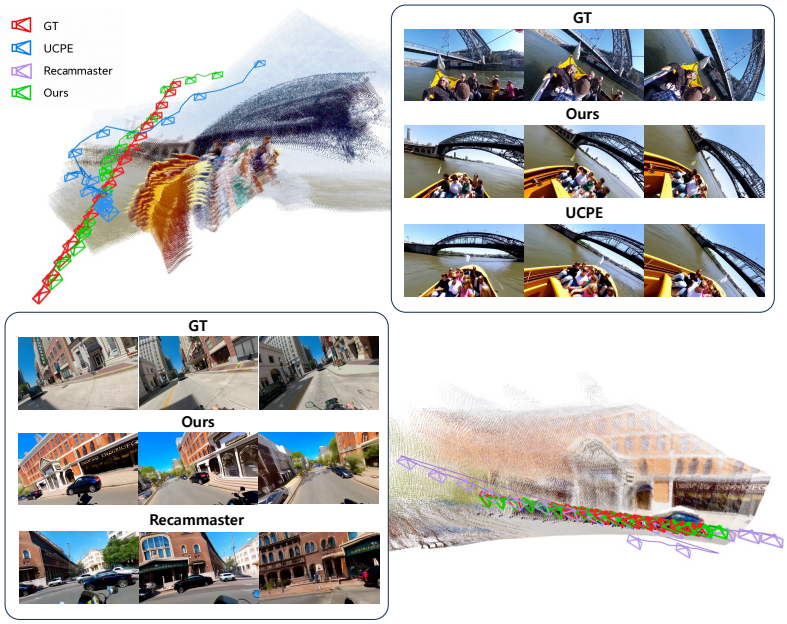

Fig. 3: Qualitative camera-control comparison. Generated videos under matched prompts and camera trajectories. The main comparison includes ReCamMaster, UCPE, and CRePE. CRePE follows the requested camera trajectory most stably under non-pinhole lenses.

4.3 Main Camera-Controlled Generation Results

Table 2 reports the comparison across lens families. CRePE improves camera-control, lens, orientation, and VBench metrics, especially where pinhole assumptions are least appropriate. On the merged set, it reduces CamMC from 21.87 to 18.15 and translation error from 17.54 to 13.78 relative to UCPE; UCPE remains slightly better on FVD (463.5 vs. 465.3). Figure 3 shows matched prompts and trajectories.

Table 2: Pinhole and non-pinhole camera-generation comparison. We report representative metrics across pose control, text/image alignment, lens distortion, orientation, and VBench quality. Non-pinhole and pinhole subsets are grouped by rows for readability. CRePE improves the representative camera-control, geometry-aware, and VBench quality metrics relative to UCPE and ReCamMaster in this comparison.

Subset	Method	Pose	Text/Image		Lens		Orientation	VBench	
		CamMC ↓	CS-I ↑	CS-T ↑	k_1 ↓	k_2 ↓	Gravity ↓	Aes. Qual. ↑	BG Cons. ↑
Non-pinhole	ReCamMaster	54.56	98.69	24.60	0.186	0.154	11.19	0.5596	0.9484
	UCPE	20.60	98.80	24.78	0.183	0.154	7.08	0.5616	0.9537
	CRePE	**19.02**	**98.81**	**24.80**	**0.157**	**0.116**	**6.27**	**0.5639**	**0.9542**
Pinhole	ReCamMaster	57.83	98.65	25.58	0.179	0.152	10.98	0.5605	0.9441
	UCPE	18.50	98.59	25.86	0.175	0.151	7.10	0.5619	0.9457
	CRePE	**15.33**	**98.65**	**25.93**	**0.153**	**0.113**	**6.13**	**0.5620**	**0.9471**

Table 3: CRePE layer placement. Applying CRePE to the middle 10 blocks provides the best overall rank and the strongest pose/lens/orientation tradeoff.

Setting	Avg. Rank	Pose	Text/Image		Video		Lens		Orientation	
		Rot. Err ↓	CS-I ↑	CS-T ↑	FID ↓	FVD ↓	k_1 ↓	k_2 ↓	Pitch ↓	Up ↓
All	3.06	4.877	99.27	25.70	136.8	1457	0.1610	0.09002	6.874	6.918
First10	2.12	5.142	99.16	**25.78**	133.9	**1438**	0.1679	0.1113	6.732	6.902
Middle10	**2.00**	**4.649**	**99.30**	25.76	136.6	1491	**0.1417**	**0.08034**	**6.612**	**6.712**
Last10	2.82	4.843	99.21	25.68	**130.3**	1506	0.1574	0.08684	6.982	6.921

4.4 Isolating Curved-Path Positional Encoding

All variants in Table 4 use the same 1K-step diffusion-only objective ($\lambda_{\text{rad}} = 0$), parameter count, and data; UniK3D is used neither in training nor inference. Interior breakpoints are deterministic evaluations of the same predicted (μ, σ), not additional learned geometry. $K = 1$ collapses the interval to a point, $K = 2$ is RayRoPE-style endpoint PE, and $K \geq 3$ can represent non-affine phase paths. The largest pose gain occurs at the transition from $K = 2$ to $K = 3$ (CamMC $14.75 \rightarrow 12.31$); $K = 5$ gives the best rank over all 19 metrics, and $K = 7$ saturates.

Table 4: Supervision-free breakpoint ablation. Avg. Rank uses 19 metrics (IS Std. excluded). $K = 3$ leads pose, while $K = 5$ gives the best overall tradeoff.

Setting	Avg. Rank ↓	CamMC ↓	Rot. ↓	Gravity ↓	CS-T ↑	FID ↓
$K = 1$ (point)	4.53	17.64	6.73	18.22	20.57	242.7
$K = 2$ (endpoint PE)	3.27	14.75	5.65	10.46	25.75	**135.3**
$K = 3$ (CRePE)	2.94	**12.31**	**4.43**	9.90	25.76	139.1
$K = 5$ (CRePE)	**1.71**	13.45	4.65	**9.69**	**25.76**	136.6
$K = 7$ (CRePE)	2.56	14.23	4.70	9.80	25.70	135.5

Why this is not a sample-count effect. For normalized log-radius $t \in [0, 1]$ and phase samples $\theta_k = \theta(t_k)$, the implemented coefficient is the average of exact per-segment integrals,

$$\bar{\rho}_K = \frac{1}{K-1} \sum_{k=1}^{K-1} \frac{e^{i\theta_{k+1}} - e^{i\theta_k}}{i(\theta_{k+1} - \theta_k)}. \tag{8}$$

If $\theta(t) = at + b$ is affine, the sum telescopes exactly to $(e^{i\theta(1)} - e^{i\theta(0)})/[i(\theta(1) - \theta(0))]$, independent of K. Thus extra subdivision helps only when the deployed phase is non-affine. Exponentiating log radius, translating between cameras, range encoding, and bounded coordinates create a shared non-affinity even for pinhole cameras; the UCM denominator adds lens-dependent transverse bending. Consistently, $K = 3$ reduces CamMC over $K = 2$ by 10.8% on pinhole and 18.6% on non-pinhole clips, and RotErr by 3.9% versus 27.2%. This supports a shared viewpoint/depth benefit with an additional non-pinhole component, rather than attributing every gain solely to UCM curvature.

Table 5: Controls and zero-shot transfer. (A) The same depth as features does not reproduce CRePE. (B) Supervision mainly stabilizes long optimization; SSI-L1 measures generated-video self-consistency. (C) Zero-shot cells are CamMC/RotErr/TransErr. Pose strides differ between B and C; compare only within panels.

A. Injection mechanism

Method	CamMC	Rot.	Trans.	FVD
UCPE	21.87	6.66	17.54	**463.5**
Depth-Conc.	23.21	7.48	18.11	473.0
CRePE	**18.15**	**6.33**	**13.78**	465.3

B. Training stabilizer

Metric	1K no	1K sup	10K no	10K sup
CamMC ↓	**52.02**	56.27	22.19	**18.15**
RotErr ↓	**18.00**	18.21	8.11	**6.33**
SSI-L1 ↓	0.129	**0.117**	0.133	**0.087**

C. Zero-shot pose

Method	KITTI-360	RealCam-Vid
UCPE	22.32/16.20/9.34	12.17/7.02/7.00
CRePE (no-sup)	18.18/14.01/**5.15**	12.31/6.68/7.22
CRePE	**17.01**/**12.70**/5.59	**9.55**/**5.79**/**4.89**

4.5 Supervision, Robustness, and Zero-Shot Transfer

Geometry as phase, not extra information. Depth-Concat predicts the same pseudo radial signal under the same loss and 10K-step schedule as CRePE, then concatenates a projected feature to every token (+6M parameters). It converges to comparable quality (FVD 473.0 vs. 465.3), but performs below even depth-free UCPE on pose (Table 5A). The advantage therefore comes from making radial distance a query-dependent phase under relative camera motion, not merely providing monocular geometry.

Supervision anchors long training. At 1K steps, adding $\mathcal{L}_{\text{rad}}$ provides no camera-control advantage; at 10K, the unsupervised head's SSI-L1 degrades while the supervised head improves by 0.047, and the CamMC ordering reverses (Table 5B). Retraining on the pinhole subset with UniDepthV2 supervision [18] is comparable to UniK3D (CamMC 15.41 vs. 15.23; RotErr 5.72 vs. 5.89), showing that the anchor is not teacher-specific.

Noisy dynamic regions. On 50 PanShot clips, moving-object masks [7] identify tokens where monocular labels are least reliable. Dynamic tokens disagree more with the pseudo target (log-MAE 0.294 vs. 0.208) but predict 1.51× wider radial uncertainty than static tokens, reducing confident phase commitment. Across frames, the median log-radial statistic changes by 0.0156 for CRePE versus 0.0375 for UniK3D and 0.0084 for the SLAM-consistent ViPE reference [6]. CRePE therefore filters teacher jitter rather than copying it; we do not interpret its token-level latent as metric moving-object depth.

Generalization beyond PanShot. Without tuning, CRePE improves every pose metric over UCPE on 32 real 180° fisheye KITTI-360 scenes [12] and 100 RealCam-Vid [38] scenes drawn from RealEstate10K, DL3DV, and MiraData9K (Table 5C). Even the no-supervision model transfers on KITTI-360, while supervision provides the margin on diverse pinhole videos. Zero-shot VBench Overall also favors CRePE on both sets (0.214 vs. 0.210; 0.215 vs. 0.189).

Fig. 4: Qualitative external-radial-map comparison on DAVIS. All methods receive the same UniK3D radial control (second column); black regions are invalid far-field rays, where CRePE-Mix falls back to its own prediction. CRePE-Mix follows the supplied layout and motion through the positional-encoding pathway alone, consistent with Table 6.

4.6 External Radial-Map Conditioning and Motion Transfer

CRePE exposes the ray-position variable through the positional pathway. At inference, CRePE-Mix may replace valid predicted radial values with an external map; missing values retain the prediction. This enables scene-layout intervention and source-video motion transfer (Figure 5) without a separate feature-control branch.

We quantify adherence on 33 DAVIS-2016 sequences [16]. Each method receives the same source map; geometry is re-estimated from its generated video and compared at pixel and 16×16-pooled token resolution. Table 6 shows that CRePE-Mix is best on all pooled metrics and seven of eight pixel/token comparisons, despite modifying only positional coefficients. With UniDepthV2 input maps never used for training, it remains second-best on δ_1 and Scale Jitter. On 48 valid non-pinhole PanShot scenes, CRePE also leads VACE and OmniVDiff in AbsRel/δ_1/Scale Jitter: 0.550/0.381/0.088 versus 0.833/0.371/0.095 and 0.929/0.097/0.233, respectively.

Figure 4 illustrates this comparison: CRePE-Mix reproduces the source layout and motion, while VACE introduces structure absent from the control map and OmniVDiff shows appearance drift.

4.7 Limitations

CRePE improves geometry-aware camera conditioning but does not solve every challenge. Main-table values are single-run point estimates; multi-seed uncertainty remains future work. Pose is recovered with ViPE, while radial adherence is re-estimated with fixed UniK3D, which also supplies pseudo labels. A common evaluation pipeline, supervision-free controls, and alternate-teacher results reduce but do not remove shared-estimator bias, so these proxy metrics require qualitative

Table 6: External radial-map adherence on DAVIS-2016 (pixel/token; best in bold). Scale Jitter is the median frame-to-frame change of the fitted depth scale.

Method	SILog ↓	AbsRel ↓	Mean-scale δ_1 ↑	Scale Jitter ↓
ControlVideo [37]	0.3952/0.3842	0.4018/0.3873	0.5465/0.5486	0.2739/0.2751
Ctrl-Adapter [13]	0.4942/0.4854	0.6305/0.6150	0.4437/0.4476	0.4013/0.4023
Make-Your-Video [30]	0.5942/0.5892	0.8540/0.8426	0.3140/0.3182	0.5087/0.5089
OmniVDiff [29]	0.3764/0.3701	0.3902/0.3842	0.5096/0.5133	0.3184/0.3181
VACE [8]	**0.2982**/0.2877	0.3016/0.2929	0.6220/0.6259	0.2518/0.2521
CRePE-Mix (UniDepthV2)	0.3218/0.3059	0.3307/0.3161	0.6348/0.6386	0.2174/0.2164
CRePE-Mix	0.2990/**0.2834**	**0.2719/0.2578**	**0.6400/0.6439**	**0.1781/0.1783**

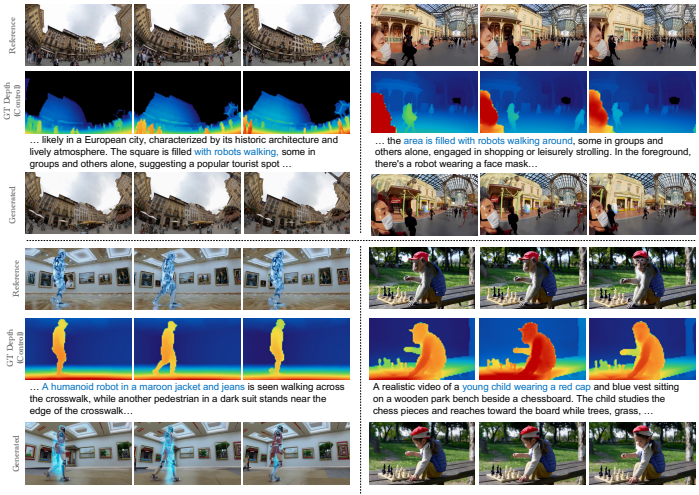

Fig. 5: External-radial-map control. CRePE-Mix uses the same positional pathway for camera-only and supplied-geometry control on PanShot (top) and in-the-wild videos (bottom).

context. Monocular supervision and per-clip scale normalization also do not establish metric moving-object depth. CRePE covers central UCM cameras, not non-central or rolling-shutter models; backbone differences preclude a state-of-the-art depth-conditioning claim, and sensor-grounded 3D/4D evaluation remains future work.

5 Conclusion

We presented CRePE, which transports log-radial uncertainty through relative UCM geometry before integrating an expected RoPE phasor. Supervision-free controls and affine-path analysis isolate non-affine integration; pseudo supervision stabilizes the radial channel. CRePE improves pinhole and non-pinhole camera control, transfers zero-shot to KITTI-360 and RealCam-Vid, and exposes the same pathway to external geometry through Radial MixForcing. It is a compact 3D state mechanism for camera-controllable video world models.

References

1. Attal, B., Huang, J.B., Zollhoefer, M., Kopf, J., Kim, C.: Learning neural light fields with ray-space embedding networks (2022), https://arxiv.org/abs/2112.01523
2. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., Zhang, D.: Recammaster: Camera-controlled generative rendering from a single video (2025), https://arxiv.org/abs/2503.11647
3. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. In: ICCV (2025)
4. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models (2024), https://arxiv.org/abs/2405.10314
5. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation (2025), https://arxiv.org/abs/2404.02101
6. Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.H., et al.: Vipe: Video pose engine for 3d geometric perception. arXiv preprint arXiv:2508.10934 (2025)
7. Huang, N., et al.: Segment any motion in videos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
8. Jiang, Z., et al.: Vace: All-in-one video creation and editing. In: International Conference on Computer Vision (2025)
9. Kong, X., Liu, S., Lyu, X., Taher, M., Qi, X., Davison, A.J.: Eschernet: A generative model for scalable view synthesis (2024), https://arxiv.org/abs/2402.03908
10. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r (2024), https://arxiv.org/abs/2406.09756
11. Li, R., Yi, B., Liu, J., Gao, H., Ma, Y., Kanazawa, A.: Cameras as relative positional encoding (2025), https://arxiv.org/abs/2507.10496
12. Liao, Y., Xie, J., Geiger, A.: Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
13. Lin, H., et al.: Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. In: International Conference on Learning Representations (2025)
14. Miyato, T., Jaeger, B., Welling, M., Geiger, A.: Gta: A geometry-aware attention mechanism for multi-view transformers (2024), https://arxiv.org/abs/2310.10375
15. Peebles, W., Xie, S.: Scalable diffusion models with transformers (2023), https://arxiv.org/abs/2212.09748
16. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2016)
17. Piccinelli, L., Sakaridis, C., Segu, M., Yang, Y.H., Li, S., Abbeloos, W., Gool, L.V.: Unik3d: Universal camera monocular 3d estimation (2025), https://arxiv.org/abs/2503.16591
18. Piccinelli, L., Yang, Y.H., Sakaridis, C.: Unidepthv2: Universal monocular metric depth estimation made simpler. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)

19. Sajjadi, M.S.M., Meyer, H., Pot, E., Bergmann, U., Greff, K., Radwan, N., Vora, S., Lucic, M., Duckworth, D., Dosovitskiy, A., Uszkoreit, J., Funkhouser, T., Tagliasacchi, A.: Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations (2022), https://arxiv.org/abs/2111.13152
20. Sitzmann, V., Rezchikov, S., Freeman, W.T., Tenenbaum, J.B., Durand, F.: Light field networks: Neural scene representations with single-evaluation rendering (2022), https://arxiv.org/abs/2106.02634
21. Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding (2023), https://arxiv.org/abs/2104.09864
22. Team, H., Wang, Z., Liu, Y., Wu, J., Gu, Z., Wang, H., Zuo, X., Huang, T., Li, W., Zhang, S., et al.: Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. arXiv preprint arXiv:2507.21809 (2025)
23. Veicht, A., Sarlin, P.E., Lindenberger, P., Pollefeys, M.: Geocalib: Learning single-image calibration with geometric optimization. In: ECCV. pp. 1–20. Springer (2024)
24. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
25. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer (2025), https://arxiv.org/abs/2503.11651
26. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy (2024), https://arxiv.org/abs/2312.14132
27. Wang, Z., Yuan, Z., Wang, X., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation (2024), https://arxiv.org/abs/2312.03641
28. Wu, Y., Jeon, M., Chang, J.H.R., Tuzel, O., Tulsiani, S.: Rayrope: Projective ray positional encoding for multi-view attention (2026), https://arxiv.org/abs/2601.15275
29. Xi, Z., et al.: Omnivdiff: Omni controllable video diffusion for generation and understanding. In: AAAI Conference on Artificial Intelligence (2026)
30. Xing, Z., et al.: Make-your-video: Customized video generation using textual and structural guidance. IEEE Transactions on Visualization and Computer Graphics (2024)
31. Yang, S., Hou, L., Huang, H., Ma, C., Wan, P., Zhang, D., Chen, X., Liao, J.: Direct-a-video: Customized video generation with user-directed camera movement and object motion. In: Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–12. SIGGRAPH '24, ACM (July 2024). https://doi.org/10.1145/3641519.3657481, http://dx.doi.org/10.1145/3641519.3657481
32. Yang, Z., Ge, W., Li, Y., Chen, J., Li, H., An, M., Kang, F., Xue, H., Xu, B., Yin, Y., et al.: Matrix-3d: Omnidirectional explorable 3d world generation. arXiv preprint arXiv:2508.08086 (2025)
33. Ye, B., Chen, B., Xu, H., Barath, D., Pollefeys, M.: Yonosplat: You only need one model for feedforward 3d gaussian splatting. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=ImRhA9xmay

34. Zhang, C., Li, B., Wei, M., Cao, Y.P., Gambardella, C.C., Phung, D., Cai, J.: Unified camera positional encoding for controlled video generation. arXiv preprint arXiv:2512.07237 (2025)
35. Zhang, J.Y., Lin, A., Kumar, M., Yang, T.H., Ramanan, D., Tulsiani, S.: Cameras as rays: Pose estimation via ray diffusion (2024), https://arxiv.org/abs/2402.14817
36. Zhang, Q., Zhai, S., Martin, M.A.B., Miao, K., Toshev, A., Susskind, J., Gu, J.: World-consistent video diffusion with explicit 3d modeling. In: CVPR. pp. 21685–21695 (2025)
37. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. In: International Conference on Learning Representations (2024)
38. Zheng, W., et al.: Realcam-vid: High-resolution video dataset with dynamic scenes and metric-scale camera movements. arXiv preprint (2025)