MEROPE: METRIC ROTARY POSITION EMBEDDING FOR CAMERA-CONTROLLED VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In camera-controlled video generation, geometry-aware positional encodings condition tokens on camera extrinsics and per-token viewing rays. Existing schemes, however, have a scale-dependent failure mode on real-world metric camera trajectories: homogeneous projective encodings can cause attention logits and feature norms to grow unbounded with physical translation baselines. We propose MeRoPE (**Me**tric **Ro**tary **P**osition **E**mbedding), a norm-preserving relative camera encoding for attention. MeRoPE encodes relative orientations between calibrated viewing rays with orthogonal rotation blocks, maps raw metric displacements into multi-frequency rotary phases, and adds a disparity-anchored correspondence prior along the epipolar arc. This design strictly preserves feature norms, bounds pre-softmax attention logits regardless of the physical translation scale, and maintains exact invariance to global rigid coordinate changes. Across a multi-dataset driving benchmark and PanShot, which cover large-baseline trajectories and diverse camera optics, respectively, MeRoPE achieves stronger camera control than prior encodings, with the lowest reported mean joint rotation–translation error between generated camera motion and conditioning poses among the compared encodings. Code will be made publicly available.

1 INTRODUCTION

Geometry-aware positional encodings (PEs) enable camera-controlled video generation by making attention sensitive to camera poses and token-level viewing geometry. However, representing real-world *metric* translation within these encodings remains challenging: raw physical displacements can destabilize attention, whereas per-trajectory normalization discards physical scale.

Existing relative camera encodings based on homogeneous projective matrices (Miyato et al., 2024; Li et al., 2025; Zhang et al., 2026a) jointly encode rotation and translation in non-orthogonal attention operators. In these formulations, physical translation enters as an unnormalized inner product, allowing attention logits and feature norms to grow unbounded with the baseline length. Over large displacements, translational terms can dominate attention scores and suppress visual content, degrading trajectory control.

This tension yields three task-driven desiderata for camera positional encodings: (A) depth-independent full metric relative-pose dependence on $T_i^{-1}T_j$, which retains both motion direction and physical translation scale without coupling translation to an assumed scene depth; (B) strict per-token factorization, which permits independent query/key transformations compatible with standard attention; and (C) norm-preserving actions, which prevent baseline-dependent amplification of logits and features. However, Theorem 1 shows that no continuous finite-dimensional encoding can satisfy all three while remaining sensitive to metric translation. Because camera control requires both metric fidelity (A) and scale stability (C), we prioritize (A+C) and relax (B) through query-camera grouping.

We therefore introduce MeRoPE (**Me**tric **Ro**tary **P**osition **E**mbedding), a norm-preserving encoding with full metric relative-pose dependence. MeRoPE constructs a minimum-rotation (MinRot) local frame for each calibrated 3D ray, represents relative ray orientation with orthogonal blocks, and maps query-frame metric displacements into multi-frequency rotary phases. Together, these operators retain raw metric camera motion while bounding attention logits across arbitrary physical baselines.

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.

Because this pose operator is scene-independent, it does not identify potential static-scene correspondences across views. MeRoPE therefore encodes correspondence hypotheses as disparity-anchored rotations along spherical epipolar arcs. Unlike CamCo (Xu et al., 2024), this design uses epipolar geometry to express relative cross-view relations without restricting which token pairs can interact. Its angular disparity anchors span a bounded spherical arc, avoiding the dataset-specific metric-depth intervals of URoPE (Xie et al., 2026a) and the learned depth estimation of RayRoPE (Wu et al., 2026).

Contributions.

- We identify baseline-dependent logit and feature-norm growth in homogeneous projective camera encodings and formalize a trilemma among full metric relative-pose dependence, strict per-token factorization, and norm preservation.
- We propose MeRoPE, which resolves this trade-off through query-camera grouping and combines a norm-preserving metric-pose operator with disparity-anchored spherical rotations for cross-view relations.
- We demonstrate improved camera control on a multi-dataset driving benchmark and camera-diverse PanShot, with MeRoPE achieving the lowest reported mean joint rotation–translation error among the compared encodings on both benchmarks.

2 Related Work

Camera conditioning pathways in video generation. Camera trajectories control viewpoint in video synthesis for applications such as autonomous driving and robotic simulation. Existing approaches implement this control through several broad pathways. *Feature fusion and adapters* integrate compact pose or ray representations into generative backbones: MotionCtrl (Wang et al., 2024) concatenates camera extrinsics with temporal features, CameraCtrl (He et al., 2025) adds multi-scale features encoded from per-pixel Plücker rays, ReCamMaster (Bai et al., 2025) additively embeds target extrinsics into transformer features, and CamCo (Xu et al., 2024) combines Plücker-conditioned feature fusion with epipolar attention. These mechanisms provide lightweight control but do not by themselves encode relative viewpoint relations in attention. *Explicit 3D guidance and rendering* instead derives structural conditions from estimated geometry: Gen3C (Ren et al., 2025) renders point-cloud caches, Lyra 2.0 (Shen et al., 2026) combines depth-based warping with geometry-aware memory, and ViewCrafter (Yu et al., 2025b) and TrajectoryCrafter (Yu et al., 2025a) condition generation on point-cloud renderings. They provide direct spatial guidance, but errors in estimated depth, pose, or scene motion can produce holes, distortions, or misaligned conditions. *Adaptive modulation* injects camera-dependent affine transformations rather than directly fusing conditions: BulletTime (Wang et al., 2026b) combines Camera-AdaLN with camera-aware positional encoding, while CompoSIA (Zhan et al., 2026) couples AdaLN-style residual ego-motion modulation with global projective attention. *Geometry-aware positional encodings* incorporate camera relations directly into attention without constructing an external 3D proxy at inference time.

Evolution of geometry-aware positional encodings. Within this family, early formulations operated on camera-level extrinsic transformations (Kong et al., 2024; Miyato et al., 2024) and projective viewing frustums (Li et al., 2025). Specifically, CaPE (Kong et al., 2024) applies block-diagonal relative pose matrices to queries and keys in multi-view attention, GTA (Miyato et al., 2024) generalizes this to group representations of $SE(3) \times SO(2) \times SO(2)$ acting on queries, keys, and values, and PRoPE (Li et al., 2025) incorporates intrinsics to model projective viewing frustums. Subsequent methods construct per-token 3D ray representations: UCPE (Zhang et al., 2026a) builds token-level ray-local rigid frames from calibrated pixel-to-ray mappings to support diverse camera optics, and RayNova (Xie et al., 2026b) applies multi-axis RoPE to Plücker coordinates in a fixed world frame. Recent efforts further incorporate depth cues: URoPE (Xie et al., 2026a) lifts tokens to discrete depth anchors and reprojects them onto the query pixel plane, RayRoPE (Wu et al., 2026) computes expected rotary phases from predicted depth segments, and CRePE (Jin et al., 2026) models curved rays with pseudo-depth expectation distributions.

Taxonomy under the trilemma and MeRoPE. Table 1 organizes camera positional encodings by three properties: (A) depth-independent full metric relative-pose dependence on $T_i^{-1}T_j$, (B) strict per-token factorization, and (C) norm-preserving attention actions. Homogeneous and projective encodings (CaPE, GTA, PRoPE, UCPE) preserve (A+B) but violate (C), as non-orthogonal translation blocks allow attention logits to grow without bound with physical baselines. World-frame ray models such as RayNova and direction-only encodings such as ViewRope (Xiang et al., 2026) preserve (B+C) but sacrifice (A): RayNova depends on the chosen global coordinate frame, whereas ViewRope omits metric translation. Query-frame reprojection approaches (Xie et al., 2026a; Wu et al., 2026; Jin et al., 2026) relax (B) and do not satisfy (A), because their attention relations are defined in pixel coordinates. Under central projection, jointly scaling camera translation and scene depth leaves projected coordinates unchanged, so pixel-space relations identify translation only up to a depth scale—the same projective ambiguity underlying monocular scale ambiguity. Fixed or predicted depths select a scale but do not make the pixel-space relation intrinsically metric. URoPE preserves (C), while the expected phasors used by RayRoPE and CRePE are not strictly norm-preserving. MeRoPE retains (A+C) by relaxing (B) via query-camera grouping; Section 3 verifies (A+C), and Theorem 1 establishes why (B) must be relaxed.

Table 1: Taxonomy under three properties: (A) depth-independent full metric relative pose, (B) strict per-token factorization, and (C) norm-preserving attention action.

Method	(A)	(B)	(C)
CaPE (Kong et al., 2024)	**Yes**	**Yes**	No
GTA (Miyato et al., 2024)	**Yes**	**Yes**	No
PRoPE (Li et al., 2025)	**Yes**	**Yes**	No
UCPE (Zhang et al., 2026a)	**Yes**	**Yes**	No
RayNova (Xie et al., 2026b)	No	**Yes**	**Yes**
ViewRope (Xiang et al., 2026)	No	**Yes**	**Yes**
URoPE (Xie et al., 2026a)	No	No	**Yes**
RayRoPE (Wu et al., 2026)	No	No	No
CRePE (Jin et al., 2026)	No	No	No
MeRoPE (Ours)	**Yes**	No	**Yes**

3 METHOD

This section presents MeRoPE (**Me**tric **Ro**tary **P**osition **E**mbedding), a norm-preserving relative camera positional encoding for video generation. Section 3.1 formalizes geometry-aware attention and analyzes the scale-dependent instability of homogeneous relative-pose operators. Section 3.2 encodes ray-frame rotation and raw metric translation with orthogonal blocks, while Section 3.3 introduces disparity-anchored spherical rotations for potential static-scene correspondences.

3.1 GEOMETRY-AWARE ATTENTION AND METRIC-SCALE LIMITATION

We formalize how camera positional encodings act within a geometry-aware attention layer. Let $a = (i, p)$ denote patch p from query camera i, and let $b = (j, q)$ denote patch q from key camera j, with corresponding query, key, and value vectors $\mathbf{q}_a, \mathbf{k}_b, \mathbf{v}_b \in \mathbb{R}^d$. Geometry-aware attention modulates query-key comparison and value aggregation through a pairwise transformation operator $\mathcal{U}_{ab} \in \mathbb{R}^{d\times d}$:

$$s_{ab} = \frac{\mathbf{q}_a^\top \mathcal{U}_{ab}\mathbf{k}_b}{\sqrt{d}}, \qquad \alpha_{ab} = \frac{\exp(s_{ab})}{\sum_{b'}\exp(s_{ab'})}, \qquad \mathbf{y}_a = \sum_b \alpha_{ab}\mathcal{U}_{ab}\mathbf{v}_b. \tag{1}$$

In GTA (Miyato et al., 2024), the camera component of $\mathcal{U}_{ab}$ is a direct sum $\bigoplus \mathcal{U}_{ij}^{\text{cam}}$ of 4×4 homogeneous relative-pose blocks:

$$R_{ij} := R_i^\top R_j, \qquad \Delta\mathbf{o}_{j|i} := R_i^\top(\mathbf{o}_j - \mathbf{o}_i), \qquad \mathcal{U}_{ij}^{\text{cam}} := \begin{bmatrix} R_{ij} & \Delta\mathbf{o}_{j|i} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \tag{2}$$

where $R_i \in \mathrm{SO}(3)$ and $\mathbf{o}_i \in \mathbb{R}^3$ denote the rotation and optical center of camera i. Decomposing the 4D feature segments of $\mathbf{q}_a$ and $\mathbf{k}_b$ into 3D vectors and scalar coordinates, $[\mathbf{u}^q; h_q]$ and $[\mathbf{u}^k; h_k]$, the resulting attention score contribution is:

$$\begin{bmatrix}\mathbf{u}^q \\ h_q\end{bmatrix}^\top \mathcal{U}_{ij}^{\text{cam}} \begin{bmatrix}\mathbf{u}^k \\ h_k\end{bmatrix} = (\mathbf{u}^q)^\top R_{ij}\mathbf{u}^k + h_k(\mathbf{u}^q)^\top \Delta\mathbf{o}_{j|i} + h_q h_k. \tag{3}$$

By the Cauchy–Schwarz inequality, the magnitude of the second term is bounded by $|h_k|\,\|\mathbf{u}^q\|_2\,\|\mathbf{o}_j - \mathbf{o}_i\|_2$, a bound proportional to the physical translation baseline. In large-baseline

settings, this translational term can dominate the attention logits, suppressing both rotational geometry and semantic visual content. Similarly, the aligned value-feature norm can grow without bound with the baseline when $h_v \neq 0$:

$$\left\| \mathcal{U}_{ij}^{\text{cam}} \begin{bmatrix} \mathbf{u}^v \\ h_v \end{bmatrix} \right\|_2^2 = \left\| R_{ij}\mathbf{u}^v + h_v \Delta \mathbf{o}_{j|i} \right\|_2^2 + h_v^2. \tag{4}$$

Figure 1 empirically illustrates this limitation on a trained baseline model. Evaluating temporal attention allocations along a forward-driving trajectory, we measure the cross-frame attention mass from the center query token of the final frame across all key frames. UCPE assigns an increasing share of attention to distant early frames as the translation baseline grows, reducing the mass on adjacent frames despite their greater cross-view co-visibility. MeRoPE bounds the positional contribution and avoids this baseline-driven shift.

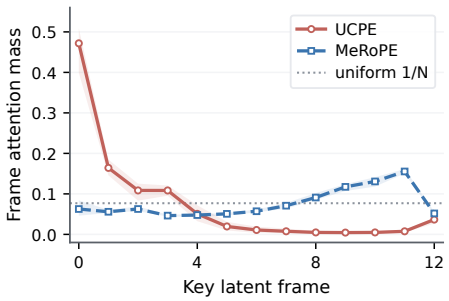

Figure 1: **Temporal attention under large baselines.** Cross-frame attention mass from the center query token of the final frame across all preceding key frames along a straight trajectory. As the baseline grows, UCPE (red) allocates more mass to distant early frames, whereas MeRoPE (blue) avoids this shift.

3.2 Orthogonal Metric Pose Encoding

The intrinsic relative camera pose $(R_i^\top R_j, \Delta \mathbf{o}_{j|i})$ consists of relative rotation and metric displacement. To prevent the norm and logit explosion of homogeneous matrices, we decompose relative pose into two decoupled orthogonal operators: a ray-local minimum-rotation (MinRot) block $\mathcal{U}_{ab}^{\text{rot}}$ and a query-frame metric translation RoPE block $\mathcal{U}_{ij}^{\text{trans}}$.

Ray-local coordinate framing via MinRot. Each image patch p from camera i is mapped to a calibrated unit viewing ray $\mathbf{d}_{i,p}^c \in \mathbb{S}^2$ in the camera coordinate frame, providing a camera-model-agnostic representation across heterogeneous optics. To make attention sensitive to individual viewing directions rather than whole-camera poses alone, each token ray is assigned a local coordinate frame ($A_{i,p}$ and $A_{j,q}$ in Figure 2). We construct this ray-local rotation matrix $A_{i,p} \in \text{SO}(3)$ via the Rodrigues minimum-rotation formula, carrying the optical axis $\mathbf{e}_z$ directly onto $\mathbf{d}_{i,p}^c$:

$$A_{i,p} = I + [\mathbf{v}_{i,p}]_\times + \frac{[\mathbf{v}_{i,p}]_\times^2}{1 + c_{i,p}}, \quad \text{where } \mathbf{v}_{i,p} = \mathbf{e}_z \times \mathbf{d}_{i,p}^c,\ c_{i,p} = \mathbf{e}_z^\top \mathbf{d}_{i,p}^c. \tag{5}$$

Unlike UCPE (Zhang et al., 2026a), whose cross-product frame degenerates when a token's vertical ray angle reaches $\pm 90°$ (equivalently, as the vertical field of view approaches $180°$), Eq. (5) is undefined only at the backward direction $\mathbf{d}_{i,p}^c = -\mathbf{e}_z$ ($180°$), outside the observed token-ray domain. Transforming ray p from camera i into world coordinates yields the ray-local frame $R_{i,p}^{\text{ray}} = R_i A_{i,p}$. The pairwise relative rotation between query ray (i,p) and key ray (j,q) is then $A_{i,p}^\top R_i^\top R_j A_{j,q}$. In attention, this relative rotation is embedded as a direct sum over m repeated 3D rotation blocks:

$$\mathcal{U}_{ab}^{\text{rot}} := \bigoplus_{k=1}^{m} \left(A_{i,p}^\top R_i^\top R_j A_{j,q} \right), \tag{6}$$

which strictly preserves feature norms while capturing the 3D relative orientation between rays.

Query-frame metric translation RoPE. To encode the query-frame metric displacement $\Delta \mathbf{o}_{j|i} = R_i^\top (\mathbf{o}_j - \mathbf{o}_i)$ illustrated in Figure 2 without baseline-dependent logit growth, we map each translation coordinate into multi-frequency rotary phases. Following RoPE frequency construction and position-interpolation principles (Su et al., 2024; Chen et al., 2023), given predefined minimum and maximum wavelengths $[\lambda_{\min,c}, \lambda_{\max,c}]$, we construct K logarithmically spaced wavelengths and their corresponding angular frequencies:

$$\omega_{c,k} = \frac{2\pi}{\lambda_{c,k}}, \qquad \lambda_{c,k} = \lambda_{\min,c} \left(\frac{\lambda_{\max,c}}{\lambda_{\min,c}} \right)^{k/(K-1)}, \qquad k = 0, \ldots, K-1. \tag{7}$$

For each Cartesian axis $c \in \{x, y, z\}$, the 1D displacement $[\Delta\mathbf{o}_{j|i}]_c$ rotates a 2D feature subspace by angle $\omega_{c,k}[\Delta\mathbf{o}_{j|i}]_c$. The translation operator is then the direct sum over all 3 axes and K frequencies:

$$\mathcal{U}_{ij}^{\text{trans}} := \bigoplus_{c\in\{x,y,z\}} \bigoplus_{k=0}^{K-1} \text{Rot}\left(\omega_{c,k}[\Delta\mathbf{o}_{j|i}]_c\right), \qquad \text{where } \text{Rot}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \tag{8}$$

which is strictly orthogonal, preserving feature norms and preventing attention logits from scaling with the physical translation baseline.

3.3 Disparity-Anchored Spherical Encoding

Potential static-scene correspondences add geometric guidance beyond the scene-independent relative camera framing provided by the orthogonal metric pose operator. Existing methods either restrict attention strictly to 2D pixel epipolar lines (Xu et al., 2024) or project metric depth anchors back into the query image plane (Xie et al., 2026a). However, pixel-plane reprojections break the camera-model-agnostic ray interface, and unbounded metric depth ranges require heuristic, dataset-specific anchor choices.

We instead construct correspondence anchors in spherical ray space using bounded angular disparity fractions (Figure 2). Let $\mathbf{u}_\infty = R_i^\top R_j \mathbf{d}_{j,q}^c$ be the key ray direction in the query camera frame; for a nonzero baseline, let $\mathbf{e} = \Delta\mathbf{o}_{j|i}/\|\Delta\mathbf{o}_{j|i}\|_2$ be the epipole unit vector. As 3D scene depth varies from ∞ to 0, points along key ray (j, q) sweep a great-circle arc on the unit sphere connecting $\mathbf{u}_\infty$ and $\mathbf{e}$. We sample L discrete disparity anchors along this epipolar arc at angular fractions $\rho_\ell \in [0, 1]$:

$$\begin{aligned} \mathbf{u}_{i,\beta_\ell} &= \cos(\rho_\ell\beta_{\max})\,\mathbf{u}_\infty + \sin(\rho_\ell\beta_{\max})\,\hat{\mathbf{t}}, \\ \text{where } \hat{\mathbf{t}} &= \frac{\mathbf{e} - (\mathbf{e}^\top\mathbf{u}_\infty)\mathbf{u}_\infty}{\|\mathbf{e} - (\mathbf{e}^\top\mathbf{u}_\infty)\mathbf{u}_\infty\|_2}, \\ \beta_{\max} &= \arccos(\mathbf{e}^\top\mathbf{u}_\infty). \end{aligned} \tag{9}$$

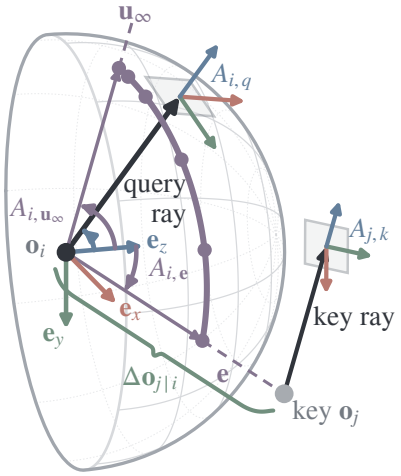

Figure 2: **Geometric construction.** Blue marks the query/key ray-local frames $A_{i,q}$ and $A_{j,k}$ used by rotation PE; green marks the query-frame metric displacement $\Delta\mathbf{o}_{j|i}$ used by translation PE; purple marks disparity PE, whose great-circle arc runs from the transformed key-ray direction $\mathbf{u}_\infty$ to the epipole $\mathbf{e}$, with endpoint frames $A_{i,\mathbf{u}_\infty}$ and $A_{i,\mathbf{e}}$ and sampled anchor directions $\mathbf{u}_{i,\beta_\ell}$ in between.

For each anchor ℓ, the Rodrigues rotation A_{i,β_ℓ} carries the query-camera optical axis $\mathbf{e}_z$ onto the query-frame direction $\mathbf{u}_{i,\beta_\ell}$. The disparity encoding computes the relative rotation $A_{i,p}^\top A_{i,\beta_\ell}$ from query ray $\mathbf{d}_{i,p}^c$ to candidate direction $\mathbf{u}_{i,\beta_\ell}$, repeated across n coordinate triplets:

$$\mathcal{U}_{(i,p),(j,q)}^{\text{disp}} := \bigoplus_{\ell=1}^{L} \underbrace{\bigoplus_{k=1}^{n} \left(A_{i,p}^\top A_{i,\beta_\ell}\right)}_{n \text{ times}}. \tag{10}$$

These anchors act as fixed geometric knots rather than predictions of mutually exclusive correspondences. Together, their rotation blocks encode the query ray relative to several fixed directions along the continuous arc, while leaving the network to determine from visual content whether a valid 3D correspondence exists.

Degenerate configurations make the construction in Eq. (9) undefined: the epipole $\mathbf{e}$ is undefined when the camera centers coincide, while the great-circle direction $\hat{\mathbf{t}}$ is undefined when $\mathbf{e}$ is parallel or antiparallel to $\mathbf{u}_\infty$. In either case, we collapse all anchor directions to $\mathbf{u}_\infty$ to avoid division by zero.

The complete pairwise MeRoPE operator $\mathcal{U}_{ab}$ is assembled by the block-diagonal direct sum:

$$\mathcal{U}_{ab} = \mathcal{U}_{(i,p),(j,q)}^{\text{disp}} \oplus \mathcal{U}_{(i,p),(j,q)}^{\text{rot}} \oplus \mathcal{U}_{ij}^{\text{trans}} \oplus \mathcal{U}^{\text{native}}, \tag{11}$$

where $\mathcal{U}^{\text{native}}$ retains the backbone's native spatio-temporal RoPE over the temporal and image-plane (x, y) coordinate bands (Figure 3). Thus, MeRoPE satisfies (A+C): it uses the full metric $T_i^{-1}T_j$ without scale normalization, while $\mathcal{U}_{ab}^{\top}\mathcal{U}_{ab} = I$ ensures norm preservation and the baseline-independent bound $|s_{ab}| \leq \|\mathbf{q}_a\|_2\|\mathbf{k}_b\|_2/\sqrt{d}$ for fixed query and key features.

Integration into Diffusion Transformers. Following the adapter scheme in UCPE (Zhang et al., 2026a), we integrate MeRoPE into pretrained video Diffusion Transformers through a lightweight parallel attention branch. The base model retains its original self-attention and spatio-temporal representations, while the parallel branch evaluates camera-conditioned attention modulated by $\mathcal{U}_{ab}$. The resulting features are mapped back through a zero-initialized linear projection and added residually to the backbone stream, so the pretrained backbone features remain unchanged at initialization.

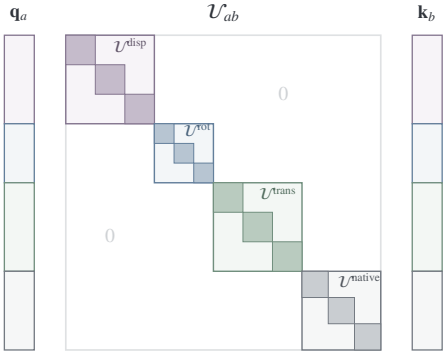

Figure 3: **Block-diagonal structure of $\mathcal{U}_{ab}$.** Pairwise attention query-key modulation $\mathbf{q}_a^{\top}\mathcal{U}_{ab}\mathbf{k}_b$, assembled as the orthogonal direct sum of disparity MinRot $\mathcal{U}^{\text{disp}}$, ray-local MinRot $\mathcal{U}^{\text{rot}}$, translation RoPE $\mathcal{U}^{\text{trans}}$, and native RoPE $\mathcal{U}^{\text{native}}$.

4 Experiments

We evaluate MeRoPE on two benchmarks: large-baseline driving video generation on a multi-dataset benchmark (Section 4.2) and multi-lens camera control on PanShot (Zhang et al., 2026a;b) (Section 4.3).

4.1 Experimental Setup

Datasets and training recipe. For driving video generation, training data combines front-camera clips from nuScenes (Caesar et al., 2020), Argoverse 2 (AV2) (Wilson et al., 2021), and Waymo (Sun et al., 2020) with synthetic clips rendered by World Engine (Li et al., 2026b), totaling about 0.75 million training clips. Candidate trajectories are grouped using k-means and filtered by non-maximum suppression (NMS) with a Fréchet distance threshold of 1 m to obtain representative motion prototypes (Figure 4). During training, the sampling schedule balances the nuScenes, AV2, and Waymo camera families and their motion distributions. All models are trained at 288×512 (nuScenes), 416×320 (AV2), or 320×480 (Waymo) with 49 frames, conditioned on the first 9 RGB frames, commanded camera trajectory parameters, and text prompts. Training uses the AdamW optimizer in bfloat16 mixed precision on $64 \times$ NVIDIA H20 GPUs for 20k iterations with a global batch size of 64 (details in Appendix B.1).

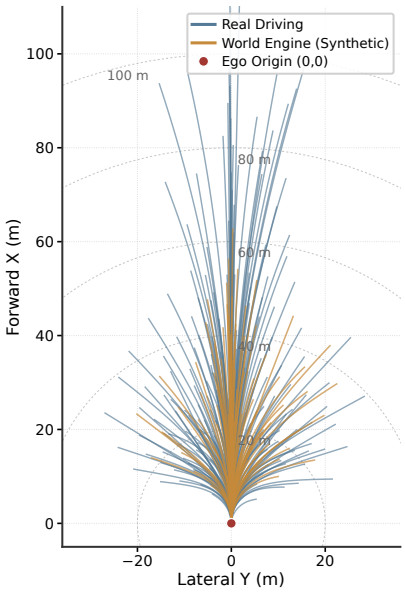

Figure 4: **Trajectory prototypes.** Representative motion patterns from real-driving datasets (blue) and World Engine (orange).

Evaluation protocols. We evaluate on 128 held-out, scene-disjoint driving sequences from nuScenes, AV2, and Waymo. Each base trajectory is augmented with left-turn and right-turn variations, yielding 384 test trajectories in total. The lateral offset is 10% of the path length, providing a geometric control stress test. Following standard visual odometry evaluation, camera poses are extracted from generated frames using VGGT-Ω (Wang et al., 2026a). We report geodesic rotation error (rot°), scale-aligned relative translation error (tr%), pairwise joint-pose AUC (AUC@3, AUC@10), and shared-scale camera motion consistency (CamMC). Translation error and CamMC use one nonnegative least-squares scale per video to account for the scale ambiguity of monocular pose recovery. Visual quality is evaluated separately using FID over 17,280 predicted frames and FVD over 432 videos (144 scenes per source), excluding the conditioning frames. Appendix B.2 defines all metrics mathematically.

Table 2: **Camera control on the mixed driving benchmark.** Methods are ordered by CamMC; best in **bold**, second-best underlined.

Method	rot°↓	tr%↓	AUC@3↑	AUC@10↑	CamMC↓	FID↓	FVD↓
No camera PE	5.424	4.880	26.872	54.432	7.091	9.257	99.932
CameraCtrl-style (He et al., 2025)	1.202	1.881	56.028	83.574	1.788	8.123	85.370
URoPE (Xie et al., 2026a)	1.144	1.875	61.320	86.196	1.718	8.271	84.638
RayRoPE (Wu et al., 2026)	1.164	**1.609**	62.989	86.655	1.662	8.450	86.892
PRoPE (Li et al., 2025)	0.947	2.087	62.150	86.216	1.627	8.559	84.623
GTA (Miyato et al., 2024)	0.942	2.056	60.354	85.430	1.612	8.172	82.540
UCPE (Zhang et al., 2026a)	<u>0.896</u>	1.920	63.581	86.533	1.508	<u>8.072</u>	**82.144**
RayNova (Xie et al., 2026b)	**0.877**	1.727	**67.499**	**88.070**	<u>1.423</u>	8.200	<u>82.521</u>
MeRoPE	0.898	<u>1.662</u>	<u>66.595</u>	<u>87.884</u>	**1.415**	**7.948**	82.979

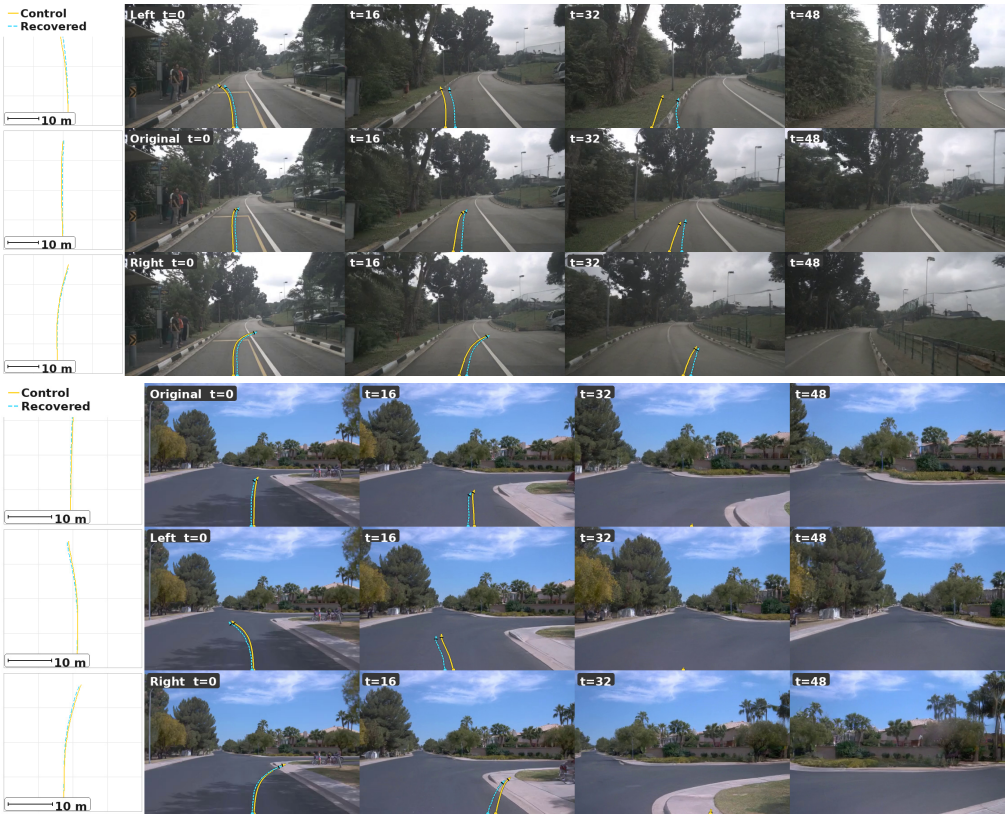

Figure 5: **Camera control in multi-source driving scenes.** A nuScenes scene (top) and a Waymo scene (bottom), each under original, left, and right commands. Each row shows the commanded (solid) and VGGT-Ω-recovered (dashed) paths, followed by frames at $t \in \{0, 16, 32, 48\}$.

For every driving configuration, both pose and quality metrics are averaged over three evaluations with paired diffusion seeds. We assess training variation through three independent runs of the default configuration (Appendix C.1) and report VGGT-Ω errors on real videos as a pose-estimation reference (Appendix C.2).

Matched-backbone baselines. For a matched-backbone comparison, all methods share the same Wan2.2 TI2V-5B backbone (Team Wan et al., 2025; Wan Team, 2025), initialization, training mixture, optimization schedule, and attention head dimension ($d = 128$). All camera-conditioned methods are retrained with the same 15-block alternating injection schedule. GTA, PRoPE, and UCPE use a common fixed global translation scale estimated solely from training-set motion statistics. This fixed rescaling preserves relative metric magnitudes across trajectories. We compare against: (i) the backbone without camera PE, (ii) GTA (Miyato et al., 2024), (iii) PRoPE (Li et al.,

Table 3: **Camera-encoding components on mixed driving scenes.** Removed blocks are replaced by identity, retaining the other blocks and $d = 128$. Best in **bold**; second-best underlined.

	Components			Overall Metrics				
Method	camera rot.	metric trans.	disparity anchor	rot°↓	tr%↓	AUC@3↑	AUC@10↑	CamMC↓
MeRoPE	✓	✓	✓	0.898	<u>1.662</u>	**66.595**	**87.884**	**1.415**
w/o disparity	✓	✓		0.912	1.744	62.195	86.143	<u>1.456</u>
w/o camera rot.		✓	✓	1.096	**1.593**	64.849	87.412	1.583
w/o metric trans.	✓		✓	<u>0.891</u>	3.063	<u>66.025</u>	<u>87.610</u>	1.988
Disparity only			✓	1.191	3.079	60.000	85.691	2.232
Rotation only	✓			**0.867**	3.362	60.961	85.325	2.094
Translation only		✓		1.494	1.881	52.543	82.029	2.061

2025), (iv) UCPE (Zhang et al., 2026a), (v) RayNova (Xie et al., 2026b), (vi) RayRoPE (Wu et al., 2026), (vii) URoPE (Xie et al., 2026a), and (viii) a CameraCtrl-style baseline (He et al., 2025), adapted to the DiT backbone by adding ray features before the query and key projections following Lyra 2.0 (Shen et al., 2026).

4.2 Pose-Controlled Video Generation on Driving Benchmarks

Main comparison on driving benchmarks. In Table 2, CamMC is our primary joint rotation–translation metric. Its observed training-run range is 0.043: MeRoPE's advantage over every baseline except RayNova exceeds this observed variation. MeRoPE improves translation control over GTA, PRoPE, and UCPE while matching or improving rotation accuracy. This pattern is consistent with the translation-control motivation of our norm-preserving, multi-frequency encoding, further examined in the component and frequency ablations. RayNova achieves comparable CamMC and rotation accuracy, with higher directional AUC scores on driving scenes; on PanShot, MeRoPE achieves lower TransErr and CamMC under both conditioning settings (Table 4). Visual quality is comparable among MeRoPE, RayNova, and UCPE within the observed FID/FVD training variation. Figure 5 compares the commanded and recovered paths under three camera controls together with the corresponding generated frames. Figure 9 provides matched-backbone qualitative comparisons on nuScenes, AV2, and Waymo.

Ablation of encoding components. Table 3 shows complementary roles for the three encoding components. Removing metric translation nearly doubles translation error despite comparable rotation accuracy and directional AUC, supporting the importance of an explicit translation positional encoding. Removing camera rotation slightly improves translation error but worsens rotation accuracy and joint control. Removing disparity reduces directional AUC, indicating that disparity anchors provide complementary geometric information. The full encoding achieves the lowest CamMC, balancing rotation and translation control.

Appendix C.3 presents further ablations of adapter capacity, injection strategies, and layer-wise modulation.

Appendix C.5 explores real-world grounding by incorporating retrieved historical images through cross-attention or joint self-attention with video tokens.

4.3 Camera-Diverse Evaluation on PanShot

To complement the multi-dataset driving benchmark, we evaluate MeRoPE on the PanShot benchmark (Zhang et al., 2026a;b), which spans pinhole, wide-angle, and fisheye lenses under the unified camera model. Following the matched protocol, we retrain every compared method from the same Wan2.1-T2V-1.3B backbone for 10k steps on $8\times$NVIDIA H20 GPUs with identical spatial attention adapters (compression ratio $C = 8$, one 192-dimensional head) and evaluate them on the official jitter-filtered test split (272 clips). We compare camera-conditioning poses with near-depth normalization (a) enabled or (b) disabled; the latter preserves raw physical displacements during training and inference. Both TransErr and CamMC follow the shared-scale definitions in Appendix B.2. Poses are extracted using VGGT-Ω (Wang et al., 2026a). Component ablations replace removed blocks by identity at fixed head width.

Table 4: **PanShot results.** Lower is better; best/second-best among compared methods are bold/underlined.

(a) Conditioning norm. on

Method	RotErr↓	TransErr↓	CamMC↓
PRoPE	5.10	16.03	18.84
GTA	5.14	15.68	18.60
RayRoPE	6.91	12.31	17.96
UCPE	4.22	12.30	14.78
RayNova	<u>3.65</u>	12.03	14.20
URoPE	**3.53**	<u>10.96</u>	<u>13.16</u>
MeRoPE	4.57	**9.63**	**12.67**
w/o disparity	5.00	13.38	16.44
w/o camera rot.	6.78	9.80	14.99
w/o metric trans.	3.92	13.10	15.06

(b) Conditioning norm. off

Method	RotErr↓	TransErr↓	CamMC↓
GTA	5.30	18.05	20.87
PRoPE	5.11	18.14	20.84
RayRoPE	6.99	<u>12.42</u>	18.01
UCPE	4.40	14.40	16.75
RayNova	<u>4.11</u>	13.49	15.98
URoPE	**3.71**	12.97	<u>15.17</u>
MeRoPE	4.82	**10.86**	**13.92**
w/o disparity	4.80	13.67	16.37
w/o camera rot.	6.48	10.37	15.05
w/o metric trans.	3.97	13.06	15.10

Table 4 shows that MeRoPE achieves the lowest TransErr and CamMC under both conditioning settings, while URoPE achieves the lowest RotErr. Component ablations exhibit the same complementary roles observed on driving scenes: removing camera rotation worsens rotation accuracy, and removing metric translation worsens translation accuracy. Removing disparity increases both TransErr and CamMC in both settings. Together, these results support the contribution of all three components to joint camera control on the camera-diverse benchmark. Figure 8 in the appendix shows generated sequences across varying fields of view and lens distortions.

5 Conclusion

We introduced MeRoPE, a norm-preserving relative camera encoding for camera-controlled video generation under real-world metric motion. Our analysis identifies baseline-dependent logit and feature-norm growth in homogeneous projective encodings and formalizes a three-way trade-off: retaining the complete metric relative pose, factorizing strictly per token, and preserving feature norms cannot be achieved simultaneously. MeRoPE resolves this trade-off by relaxing per-token factorization through query-camera grouping, representing relative ray orientation and raw metric translation with orthogonal blocks, and adding disparity-anchored spherical rotations as geometric knots for cross-view relations.

Across driving scenes and diverse camera optics, MeRoPE achieves the lowest reported mean CamMC among the compared encodings. These results support norm-preserving metric camera encoding as a stable basis for camera control across physical baselines and camera optics. Query-camera grouping introduces additional local camera-attention cost; reducing this overhead while retaining metric sensitivity is an important direction for future work.

AI use statement

Generative AI tools assisted with editing the manuscript for clarity, implementing research code and prior-method baselines, and developing mathematical proofs. The authors checked the relevance and accuracy of all references and citations.

For implementations of prior methods, AI agents consulted the original open-source code, multiple agents cross-checked reproduction fidelity, and the authors tested the resulting code. Hyperparameters were adapted to our benchmark.

The authors manually verified the proofs and reviewed all AI-assisted text and code, and take full responsibility for this paper.

References

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. ReCamMaster: Camera-controlled generative

rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14834–14844, 2025. doi: 10.1109/ICCV51701.2025.01376. URL https://arxiv.org/abs/2503.11647.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020. doi: 10.1109/CVPR42600.2020.01164. URL https://arxiv.org/abs/1903.11027.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023. URL https://arxiv.org/abs/2306.15595.

Gene Chou, Charles Herrmann, Kyle Genova, Boyang Deng, Songyou Peng, Bharath Hariharan, Jason Y. Zhang, Noah Snavely, and Philipp Henzler. CityRAG: Stepping into a city via spatially-grounded video generation. *arXiv preprint arXiv:2604.19741*, 2026. URL https://arxiv.org/abs/2604.19741.

Gerald B. Folland. *A Course in Abstract Harmonic Analysis*. CRC Press, 2nd edition, 2016.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2404.02101.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

Seonghyun Jin, Youngmin Kim, Sunwoo Park, and Jong Chul Ye. CRePE: Curved ray expectation positional encoding for unified-camera-controlled video generation. *arXiv preprint arXiv:2605.12938*, 2026. URL https://arxiv.org/abs/2605.12938.

Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J. Davison. EscherNet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9503–9513, 2024. doi: 10.1109/CVPR52733.2024.00908. URL https://arxiv.org/abs/2402.03908.

Chunyang Li, Yuanbo Yang, Jiahao Shao, Hongyu Zhou, Katja Schwarz, and Yiyi Liao. ReRoPE: Repurposing RoPE for relative camera control. In *ACM SIGGRAPH 2026 Conference Papers*, pp. 1–11, 2026a. doi: 10.1145/3799902.3811226. URL https://arxiv.org/abs/2602.08068.

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*, pp. 18020–18045, 2025. doi: 10.52202/085713-0540. URL https://arxiv.org/abs/2507.10496.

Tianyu Li, Li Chen, Caojun Wang, Haochen Liu, Kashyap Chitta, Zhenjie Yang, Yuhang Lu, Naisheng Ye, Yihang Qiu, Yufei Wang, Luoxi Zou, Jiaxin Peng, Jin Pan, Zhaoyu Su, Andrei Bursuc, Shengbo Eben Li, Andreas Geiger, Peng Su, and Hongyang Li. World Engine: Towards the era of post-training for autonomous driving. *arXiv preprint arXiv:2606.19836*, 2026b. URL https://arxiv.org/abs/2606.19836.

Takeru Miyato, Bernhard Jaeger, Max Welling, and Andreas Geiger. GTA: A geometry-aware attention mechanism for multi-view transformers. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.10375.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3C: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025. doi: 10.1109/CVPR52734.2025.00574. URL https://arxiv.org/abs/2503.03751.

Junyoung Seo, Hyunwook Choi, Minkyung Kwon, Jinhyeok Choi, Siyoon Jin, Gayoung Lee, Junho Kim, JoungBin Lee, Geonmo Gu, Dongyoon Han, Sangdoo Yun, Seungryong Kim, and Jin-Hwa Kim. Grounding world simulation models in a real-world metropolis. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2603.15583. Spotlight; Seoul World Model.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, Sanja Fidler, Jiahui Huang, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra 2.0: Explorable generative 3D worlds. *arXiv preprint arXiv:2604.13036*, 2026. URL https://arxiv.org/abs/2604.13036.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://arxiv.org/abs/2104.09864.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/1912.04838.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Wan Team. Wan2.2: Open-source video generation models. Official GitHub repository, 2025. URL https://github.com/Wan-Video/Wan2.2.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes L. Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21486–21499, 2026a.

Yiming Wang, Qihang Zhang, Shengqu Cai, Tong Wu, Jan Ackermann, Zhengfei Kuang, Yang Zheng, Frano Rajič, Siyu Tang, and Gordon Wetzstein. BulletTime: Decoupled control of time and camera pose for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18319–18330, 2026b.

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 Conference Papers*, pp. 1–11, 2024. doi: 10.1145/3641519.3657518. URL https://arxiv.org/abs/2312.03641.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://www.argoverse.org/av2.html.

Yu Wu, Minsik Jeon, Jen-Hao Rick Chang, Oncel Tuzel, and Shubham Tulsiani. RayRoPE: Projective ray positional encoding for multi-view attention. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2601.15275. Spotlight.

Chendong Xiang, Jiajun Liu, Jintao Zhang, Xiao Yang, Zhengwei Fang, Shizun Wang, Zijun Wang, Yingtian Zou, Hang Su, and Jun Zhu. Geometry-aware rotary position embedding for consistent video world model. *arXiv preprint arXiv:2602.07854*, 2026. URL https://arxiv.org/abs/2602.07854.

Yichen Xie, Depu Meng, Chensheng Peng, Yihan Hu, Quentin Herau, Masayoshi Tomizuka, and Wei Zhan. URoPE: Universal relative position embedding across geometric spaces. In *European Conference on Computer Vision*, 2026a. URL https://arxiv.org/abs/2604.18747. Spotlight.

Yichen Xie, Chensheng Peng, Mazen Abdelfattah, Yihan Hu, Jiezhi Yang, Eric Higgins, Ryan Brigden, Masayoshi Tomizuka, and Wei Zhan. RayNova: Scale-temporal autoregressive world modeling in ray space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25426–25437, 2026b.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. CamCo: Camera-controllable 3D-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. URL https://arxiv.org/abs/2406.02509.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 100–111, 2025a. doi: 10.1109/ICCV51701.2025.00017. URL https://arxiv.org/abs/2503.05638.

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*, 2025b. doi: 10.1109/TPAMI.2025.3613256. URL https://arxiv.org/abs/2409.02048.

Yifan Zhan, Zhengqing Chen, Qingjie Wang, Zhuo He, Muyao Niu, Xiaoyang Guo, Wei Yin, Weiqiang Ren, Qian Zhang, and Yinqiang Zheng. Composing driving worlds through disentangled control for adversarial scenario generation. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2603.12864.

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*, pp. 38027–38037, 2026a.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Cruz Gambardella, Dinh Phung, and Jianfei Cai. UCPE: Official implementation and PanShot dataset. GitHub repository and Hugging Face dataset, 2026b. URL https://github.com/chengzhag/UCPE. Commit d992f1807803ba99331e807e8f018ed552886afd.

A Impossibility of Unitary Relative Per-Token Encodings for Metric Translation

We show that no strictly per-token, unitary (or orthogonal) positional operator can completely capture metric relative translation. This result formalizes the trilemma introduced in Section 1 and motivates MeRoPE's use of query-camera grouping.

A.1 Problem Formulation and Hypotheses

Let $T_i = (R_i, \mathbf{o}_i) \in SE(3)$ denote the camera-to-world pose of view i, where $R_i \in \mathrm{SO}(3)$ and $\mathbf{o}_i \in \mathbb{R}^3$. The intrinsic relative pose between views i and j, invariant to any global coordinate change $G \in SE(3)$, is given by

$$T_i^{-1} T_j = \big(R_i^\top R_j,\ R_i^\top (\mathbf{o}_j - \mathbf{o}_i)\big). \tag{12}$$

We examine the existence of a matrix-valued positional embedding operator $U : SE(3) \to \mathbb{C}^{d\times d}$ (or $\mathbb{R}^{d\times d}$) satisfying four standard axioms of relative positional attention:

(H1) **Multiplicative per-token factorization:** The camera positional operator acts independently on each query and key token via its own pose:

$$\widetilde{\mathbf{q}}_i = U(T_i)\mathbf{q}_i, \qquad \widetilde{\mathbf{k}}_j = U(T_j)\mathbf{k}_j, \qquad s_{ij} = \frac{\mathbf{q}_i^* U(T_i)^* U(T_j)\mathbf{k}_j}{\sqrt{d}}. \tag{13}$$

(H2) **Exact relative-pose dependence:** The pairwise interaction depends exclusively on the relative pose $T_i^{-1}T_j$:

$$U(T_i)^*U(T_j) = U(T_i^{-1}T_j), \qquad \forall T_i, T_j \in SE(3). \tag{14}$$

(H3) **Norm preservation (unitarity):** $U(T)$ is unitary (or orthogonal in the real case), so that $\|U(T)\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ for all $\mathbf{x} \in \mathbb{C}^d$, i.e.,

$$U(T)^*U(T) = I_d, \qquad \forall T \in SE(3). \tag{15}$$

(H4) **Finite-dimensional continuity:** The feature dimension is finite and positive ($1 \le d < \infty$), and the mapping $T \mapsto U(T)$ is continuous with respect to the standard Lie group topology on $SE(3)$.

A.2 The Impossibility Theorem

Theorem 1 (Translation Blindness of Unitary Relative Per-Token Encodings)**.** *Let $U : SE(3) \to \mathrm{U}(d)$ be any continuous map satisfying hypotheses* (H1)–(H4)*. Then for every pure metric translation $\mathbf{t} \in \mathbb{R}^3$,*

$$U(I_3, \mathbf{t}) = I_d. \tag{16}$$

Consequently, for any camera pose $(R, \mathbf{t}) \in SE(3)$,

$$U(R, \mathbf{t}) = U(R, \mathbf{0}). \tag{17}$$

The positional operator U is identically trivial on metric translations, failing to distinguish any non-zero relative translation $\Delta\mathbf{o}_{j|i} = R_i^\top(\mathbf{o}_j - \mathbf{o}_i) \neq \mathbf{0}$.

Proof. **Step 1: Homomorphism deduction.** First, we show that (H2) and (H3) imply that U is a group representation of $SE(3)$, i.e., $U(gh) = U(g)U(h)$ for all $g, h \in SE(3)$:

- Setting $T_i = T_j = I$ in Eq. (14) yields $U(I)^*U(I) = U(I)$. On the other hand, Eq. (15) with $T = I$ gives $U(I)^*U(I) = I_d$. The two right-hand sides must coincide, so $U(I) = I_d$.
- Setting $T_j = I$ in Eq. (14) yields $U(T_i)^*U(I) = U(T_i^{-1})$. Since $U(I) = I_d$ from the first item, the left-hand side reduces to $U(T_i)^*$, and by unitarity ($U(T_i)^{-1} = U(T_i)^*$) we conclude $U(T_i^{-1}) = U(T_i)^* = U(T_i)^{-1}$.
- For any $g, h \in SE(3)$, set $T_i = g$ and $T_j = gh$. Substituting into Eq. (14) gives:

  $$U(g)^*U(gh) = U(g^{-1}(gh)) = U(h). \tag{18}$$

  Multiplying on the left by $U(g)$ and using $U(g)U(g)^* = I_d$, we conclude that

  $$U(gh) = U(g)U(h), \qquad \forall g, h \in SE(3). \tag{19}$$

Step 2: Commutativity of the translation subgroup. Let $U_{\mathrm{tr}}(\mathbf{t}) := U(I_3, \mathbf{t})$ denote the pure translation operator for $\mathbf{t} \in \mathbb{R}^3$. In $SE(3)$, pure translations commute: $(I_3, \mathbf{s})(I_3, \mathbf{t}) = (I_3, \mathbf{s}+\mathbf{t}) = (I_3, \mathbf{t})(I_3, \mathbf{s})$. Applying the homomorphism in Eq. (19):

$$U_{\mathrm{tr}}(\mathbf{s})U_{\mathrm{tr}}(\mathbf{t}) = U\big((I_3, \mathbf{s})(I_3, \mathbf{t})\big) = U(I_3, \mathbf{s}+\mathbf{t}) = U\big((I_3, \mathbf{t})(I_3, \mathbf{s})\big) = U_{\mathrm{tr}}(\mathbf{t})U_{\mathrm{tr}}(\mathbf{s}) = U_{\mathrm{tr}}(\mathbf{s}+\mathbf{t}). \tag{20}$$

Thus, $\{U_{\mathrm{tr}}(\mathbf{t}) : \mathbf{t} \in \mathbb{R}^3\}$ is a family of mutually commuting unitary matrices.

Step 3: Simultaneous unitary diagonalization. Every unitary matrix is normal ($U^*U = UU^* = I$). Step 2 shows that the operators $\{U_{\mathrm{tr}}(\mathbf{t}) : \mathbf{t} \in \mathbb{R}^3\}$ pairwise commute, and each is unitary by (H3), hence normal. The spectral theorem for commuting families of normal operators (Horn & Johnson, 2012, Theorem 2.5.5), which applies to families of arbitrary cardinality, then guarantees a single, $\mathbf{t}$-independent unitary change-of-basis matrix $V \in \mathrm{U}(d)$ that simultaneously diagonalizes all $U_{\mathrm{tr}}(\mathbf{t})$:

$$V^*U_{\mathrm{tr}}(\mathbf{t})V = \mathrm{diag}\big(\chi_1(\mathbf{t}), \chi_2(\mathbf{t}), \ldots, \chi_d(\mathbf{t})\big). \tag{21}$$

By (H4), for each $a \in \{1, \dots, d\}$, the diagonal entry $\chi_a : \mathbb{R}^3 \to \mathbb{C}$ is continuous. By (H3), $|\chi_a(\mathbf{t})| = 1$. It remains to show that each entry is multiplicative in $\mathbf{t}$. Conjugating Eq. (20) by V and inserting $VV^* = I_d$ between $U_{\mathrm{tr}}(\mathbf{s})$ and $U_{\mathrm{tr}}(\mathbf{t})$:

$$V^* U_{\mathrm{tr}}(\mathbf{s}+\mathbf{t}) V = \big(V^* U_{\mathrm{tr}}(\mathbf{s}) V\big)\big(V^* U_{\mathrm{tr}}(\mathbf{t}) V\big). \tag{22}$$

By Eq. (21), the left-hand side equals $\mathrm{diag}\big(\chi_1(\mathbf{s}+\mathbf{t}), \dots, \chi_d(\mathbf{s}+\mathbf{t})\big)$, and the right-hand side equals $\mathrm{diag}\big(\chi_1(\mathbf{s})\chi_1(\mathbf{t}), \dots, \chi_d(\mathbf{s})\chi_d(\mathbf{t})\big)$. Comparing diagonal entries, each entry satisfies the multiplicative Cauchy character equation:

$$\chi_a(\mathbf{s}+\mathbf{t}) = \chi_a(\mathbf{s})\chi_a(\mathbf{t}), \qquad \forall \mathbf{s}, \mathbf{t} \in \mathbb{R}^3. \tag{23}$$

Step 4: Character form and finite spatial spectrum. Fixing an index a and dropping it, we classify the continuous maps $\chi : (\mathbb{R}^3, +) \to \mathbb{T}$, where $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$, satisfying the multiplicative character equation

$$\chi(\mathbf{s}+\mathbf{t}) = \chi(\mathbf{s})\chi(\mathbf{t}), \qquad |\chi| = 1, \tag{24}$$

whose only solutions are spatial plane waves; this is the classical classification of continuous characters of $(\mathbb{R}^3, +)$ (Folland, 2016, Theorem 4.6 and Corollary 4.8):

$$\chi_a(\mathbf{t}) = \exp(i\boldsymbol{\omega}_a^\top \mathbf{t}), \qquad \text{for a unique frequency vector } \boldsymbol{\omega}_a \in \mathbb{R}^3. \tag{25}$$

This classification enables the spherical-orbit contradiction below. Consequently, in the eigenbasis of V, every translation operator takes the standard multi-frequency phase form:

$$V^* U_{\mathrm{tr}}(\mathbf{t}) V = \mathrm{diag}\left(\exp(i\boldsymbol{\omega}_1^\top \mathbf{t}), \dots, \exp(i\boldsymbol{\omega}_d^\top \mathbf{t})\right). \tag{26}$$

Because the feature dimension is finite ($d < \infty$), the active frequency spectrum

$$\Omega := \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \dots, \boldsymbol{\omega}_d\} \subset \mathbb{R}^3 \tag{27}$$

contains at most d discrete frequency vectors ($|\Omega| \le d$).

Step 5: Semi-direct product conjugation and spherical orbit contradiction.

A lemma on rotation invariance. If a nonempty finite set $\Omega \subset \mathbb{R}^3$ is $\mathrm{SO}(3)$-invariant, i.e. $R\Omega = \Omega$ for all $R \in \mathrm{SO}(3)$, then $\Omega = \{\mathbf{0}\}$. Indeed, the orbit of any non-zero $\boldsymbol{\omega} \in \Omega$ under $\mathrm{SO}(3)$ is the entire sphere $\{\mathbf{u} \in \mathbb{R}^3 : \|\mathbf{u}\|_2 = \|\boldsymbol{\omega}\|_2\}$, an uncountably infinite set; invariance forces this whole sphere into Ω, which finiteness forbids.

Let $U_{\mathrm{rot}}(R) := U(R, \mathbf{0})$ denote the pure rotation operator for $R \in \mathrm{SO}(3)$. In $SE(3)$, rotations and translations obey the semi-direct product conjugation:

$$(I_3, R\mathbf{t}) = (R, \mathbf{0})(I_3, \mathbf{t})(R, \mathbf{0})^{-1}, \qquad \forall R \in \mathrm{SO}(3),\ \mathbf{t} \in \mathbb{R}^3. \tag{28}$$

Applying the homomorphism U to Eq. (28) gives the operator-level conjugation

$$U_{\mathrm{tr}}(R\mathbf{t}) = U_{\mathrm{rot}}(R) U_{\mathrm{tr}}(\mathbf{t}) U_{\mathrm{rot}}(R)^{-1}. \tag{29}$$

Since $\mathbf{t}$ is a free variable ranging over $\mathbb{R}^3$, replacing it by $R^{-1}\mathbf{t}$ and then multiplying both sides on the right by $U_{\mathrm{rot}}(R)$ (the factors $U_{\mathrm{rot}}(R)^{-1} U_{\mathrm{rot}}(R) = I_d$ cancel) recasts the same fact in the commutation form

$$U_{\mathrm{tr}}(\mathbf{t}) U_{\mathrm{rot}}(R) = U_{\mathrm{rot}}(R) U_{\mathrm{tr}}(R^{-1}\mathbf{t}), \tag{30}$$

which is the form we use below to push $U_{\mathrm{rot}}(R)$ past $U_{\mathrm{tr}}(\mathbf{t})$ when acting on an eigenvector. For any frequency $\boldsymbol{\omega} \in \Omega$, define its joint eigenspace, and set $E_{\boldsymbol{\omega}} = \{\mathbf{0}\}$ for $\boldsymbol{\omega} \notin \Omega$:

$$E_{\boldsymbol{\omega}} = \left\{\mathbf{v} \in \mathbb{C}^d : U_{\mathrm{tr}}(\mathbf{t})\mathbf{v} = \exp(i\boldsymbol{\omega}^\top \mathbf{t})\mathbf{v},\ \forall \mathbf{t} \in \mathbb{R}^3\right\}, \quad E_{\boldsymbol{\omega}} = \{\mathbf{0}\} \text{ if } \boldsymbol{\omega} \notin \Omega. \tag{31}$$

For any non-zero $\mathbf{v} \in E_{\boldsymbol{\omega}}$, we apply Eq. (30) to evaluate the translation action on $U_{\mathrm{rot}}(R)\mathbf{v}$:

$$\begin{aligned} U_{\mathrm{tr}}(\mathbf{t})\big(U_{\mathrm{rot}}(R)\mathbf{v}\big) &= U_{\mathrm{rot}}(R) U_{\mathrm{tr}}(R^{-1}\mathbf{t})\mathbf{v} \\ &= U_{\mathrm{rot}}(R) \exp(i\boldsymbol{\omega}^\top R^{-1}\mathbf{t})\mathbf{v} \\ &= \exp(i(R\boldsymbol{\omega})^\top \mathbf{t})\big(U_{\mathrm{rot}}(R)\mathbf{v}\big). \end{aligned} \tag{32}$$

Here the second equality uses that $\mathbf{v} \in E_{\boldsymbol{\omega}}$ is an eigenvector of $U_{\mathrm{tr}}(R^{-1}\mathbf{t})$; the last uses that the (unit-modulus) eigenvalue $\exp(i\boldsymbol{\omega}^\top R^{-1}\mathbf{t})$ is a complex scalar, hence commutes with $U_{\mathrm{rot}}(R)$

and can be pulled out of the product, together with $\boldsymbol{\omega}^\top R^{-1}\mathbf{t} = (R\boldsymbol{\omega})^\top \mathbf{t}$. Thus, $U_{\mathrm{rot}}(R)\mathbf{v}$ is an eigenvector of $U_{\mathrm{tr}}(\mathbf{t})$ with spatial frequency $R\boldsymbol{\omega}$, meaning $U_{\mathrm{rot}}(R)E_{\boldsymbol{\omega}} = E_{R\boldsymbol{\omega}}$. The frequency spectrum Ω must therefore be invariant under all continuous 3D rotations:

$$R\Omega = \Omega, \qquad \forall R \in \mathrm{SO}(3). \tag{33}$$

By Eq. (33) and the finiteness $|\Omega| \leq d < \infty$ from Step 4, Ω is a finite $\mathrm{SO}(3)$-invariant subset of $\mathbb{R}^3$; the lemma therefore forces $\Omega = \{\mathbf{0}\}$.

Substituting $\boldsymbol{\omega}_a = \mathbf{0}$ back into the diagonalized form yields

$$U_{\mathrm{tr}}(\mathbf{t}) = V I_d V^* = I_d, \qquad \forall \mathbf{t} \in \mathbb{R}^3. \tag{34}$$

Finally, decomposing any pose as $(R, \mathbf{t}) = (I_3, \mathbf{t})(R, \mathbf{0})$ and applying the homomorphism yields $U(R, \mathbf{t}) = U_{\mathrm{tr}}(\mathbf{t})U_{\mathrm{rot}}(R) = U(R, \mathbf{0})$. This completes the proof. □

B Implementation and Evaluation Details

B.1 Architecture and Training Configurations

Head partition. In our primary implementation based on the Wan2.2 TI2V-5B backbone ($d = 128$), the channel budget is partitioned into four orthogonal sub-blocks: $d = d_{\mathrm{disp}} + d_{\mathrm{rot}} + d_{\mathrm{trans}} + d_{\mathrm{native}} = 36 + 36 + 24 + 32$. Specifically, $d_{\mathrm{disp}} = 36$ encodes $L = 6$ non-uniform angular disparity fractions $\rho_\ell \in \{0, 0.05, 0.15, 0.4, 0.7, 1\}$ with $n = 2$ triplets each ($6 \times 2 \times 3$); $d_{\mathrm{rot}} = 36$ allocates $m = 12$ coordinate triplets for ray-local MinRot; $d_{\mathrm{trans}} = 24$ covers 3 spatial axes over $K = 4$ log-spaced metric wavelengths ($[0.5, 50]$, $[0.5, 6]$, and $[0.5, 200]$ m for the query-frame x, y, z axes, respectively); and $d_{\mathrm{native}} = 32$ applies standard RoPE to temporal (16D) and image-plane x/y (8D each) coordinate bands.

For PanShot, the 192-dimensional head contains 36 disparity, 36 rotation, and 120 translation channels, with translation wavelengths in $[0.5, 8]$. RayNova uses time and ray phase scales of 18 and 50, respectively; moment scale 8 gives the lowest CamMC among $\{1, 2, 4, 8\}$ under both conditioning settings.

Backward disparity anchors. We evaluate anchor rotations in axis-angle form. At $\mathbf{u}_{i,\beta_\ell} = -\mathbf{e}_z$, where Eq. (5) is undefined, we choose $A_{i,\beta_\ell} = \mathrm{diag}(-1, 1, -1)$, a 180° rotation about $\mathbf{e}_y$. This preserves orthogonality, although the frame choice is not continuous at that direction.

Multi-source driving data. The training resolutions are 288×512 for nuScenes, 416×320 for AV2, and 320×480 for Waymo. Images are resized without cropping, and intrinsics are scaled accordingly. AV2 uses rectified pinhole images, while Waymo retains Brown distortion. The sampling schedule balances the nuScenes, AV2, and Waymo camera families and their motion distributions. Real clips use nuScenes and AV2 train/validation splits and Waymo training; synthetic clips use training-source assets after filtering severe static-scene defects. GTA, PRoPE, and UCPE use a shared translation scale of approximately 0.02188, estimated from the training trajectories.

Optimization details. All driving models are trained with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay 0.01) using a constant learning rate of 1×10^{-5} and no warmup. Training uses bfloat16 mixed precision on $64 \times$ NVIDIA H20 GPUs for 20k iterations with a global batch size of 64. PanShot models are trained on $8 \times$ NVIDIA H20 GPUs for 10k iterations.

B.2 Evaluation Metrics and Mathematical Formulations

Estimated camera poses are recovered from generated frames by VGGT-Ω (Wang et al., 2026a). Let T denote the number of frames in an evaluated sequence, and let $T_t = (R_t, \mathbf{o}_t) \in SE(3)$ and $\widehat{T}_t = (\widehat{R}_t, \widehat{\mathbf{o}}_t) \in SE(3)$ denote the commanded and estimated camera-to-world poses at frame $t \in \{1, \ldots, T\}$. To remove the absolute $SE(3)$ gauge, all poses are expressed relative to the first frame,

$$T_t^{\mathrm{rel}} = T_1^{-1} T_t = \big(R_1^\top R_t,\ R_1^\top(\mathbf{o}_t - \mathbf{o}_1)\big) =: \big(R_t^{\mathrm{rel}},\ \mathbf{o}_t^{\mathrm{rel}}\big), \tag{35}$$

with $\widehat{T}_t^{\mathrm{rel}} = (\widehat{R}_t^{\mathrm{rel}}, \widehat{\mathbf{o}}_t^{\mathrm{rel}})$ constructed analogously. The geodesic rotation error in degrees is

$$\angle(R, \widehat{R}) = \arccos\left(\mathrm{clamp}\left(\frac{\mathrm{tr}(R^\top \widehat{R}) - 1}{2}, -1, 1\right)\right) \times \frac{180^\circ}{\pi}. \tag{36}$$

Sequence-level metrics are averaged equally over evaluated clips; pairwise AUC pools all within-clip frame pairs.

Monocular pose recovery leaves a global scale ambiguity. For driving translation error, PanShot TransErr, and CamMC in all settings, we fit one nonnegative least-squares scale per video over the full estimated trajectory, leaving rotations unchanged:

$$s^\star = \max\left(0, \frac{\sum_{t=1}^{T} \widehat{\mathbf{o}}_t^{\mathrm{rel}\top} \mathbf{o}_t^{\mathrm{rel}}}{\sum_{t=1}^{T} \|\widehat{\mathbf{o}}_t^{\mathrm{rel}}\|_2^2}\right), \qquad \widetilde{\mathbf{o}}_t = s^\star\, \widehat{\mathbf{o}}_t^{\mathrm{rel}}. \tag{37}$$

If the denominator is at most 10^{-12}, we set $s^\star = 0$. The nonnegative constraint prevents alignment from reversing the estimated motion. For CamMC, we define a common ground-truth normalization scale and apply it to both trajectories:

$$a = \max_{1 \le \tau \le T} \|\mathbf{o}_\tau^{\mathrm{rel}}\|_2 + \epsilon_0, \qquad \mathbf{c}_t = \mathbf{o}_t^{\mathrm{rel}}/a, \qquad \widehat{\mathbf{c}}_t = s^\star \widehat{\mathbf{o}}_t^{\mathrm{rel}}/a, \tag{38}$$

Here $\epsilon_0 = 10^{-9}$. Unlike UCPE, which independently normalizes each trajectory by its maximum translation norm (Zhang et al., 2026b), we apply the same ground-truth normalization after one scale alignment per video. Both PanShot TransErr and CamMC use Eq. (38).

Rotation Error (rot°). Used on driving evaluations. The first-frame relative pose is the identity, so the mean excludes $t = 1$:

$$\mathrm{rot}^\circ = \frac{1}{T-1} \sum_{t=2}^{T} \angle\big(R_t^{\mathrm{rel}},\, \widehat{R}_t^{\mathrm{rel}}\big). \tag{39}$$

Relative Translation Error (tr%). Used on driving evaluations. After the scale alignment in Eq. (37), the mean center error is normalized by the ground-truth endpoint displacement (chord length):

$$\mathrm{tr}\% = \frac{\frac{1}{T-1} \sum_{t=2}^{T} \|\mathbf{o}_t^{\mathrm{rel}} - \widetilde{\mathbf{o}}_t\|_2}{\|\mathbf{o}_T^{\mathrm{rel}}\|_2} \times 100\%. \tag{40}$$

Pairwise Joint Pose AUC (AUC@τ). Used on driving evaluations. Let $\mathcal{P} = \{(i, j) : 1 \le i < j \le T\}$ denote within-clip frame pairs, pooled over the evaluation set. For each pair, form the relative motions $T_{i\to j} = (T_i^{\mathrm{rel}})^{-1} T_j^{\mathrm{rel}}$ and $\widehat{T}_{i\to j} = (\widehat{T}_i^{\mathrm{rel}})^{-1} \widehat{T}_j^{\mathrm{rel}}$, and define the unoriented translation-direction error

$$e_{ij}^{\mathrm{t}} = \arccos\left(\mathrm{clamp}\left(\left|\frac{\mathbf{o}_{i\to j}^\top \widehat{\mathbf{o}}_{i\to j}}{\|\mathbf{o}_{i\to j}\|_2 \|\widehat{\mathbf{o}}_{i\to j}\|_2}\right|, 0, 1\right)\right) \times \frac{180^\circ}{\pi}, \tag{41}$$

with e_{ij}^{t} set to a large sentinel when either translation vanishes. The joint pair error is $e_{ij} = \max\big(\angle(R_{i\to j}, \widehat{R}_{i\to j}),\, e_{ij}^{\mathrm{t}}\big)$. Because e_{ij}^{t} depends only on direction, this metric does not apply $s^\star$. With integer thresholds $\tau \in \{3, 10\}$ (degrees), the discrete threshold average is

$$\mathrm{AUC@}\tau = \frac{1}{\tau} \sum_{k=1}^{\tau} \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \mathbb{I}(e_{ij} < k^\circ) \times 100\%. \tag{42}$$

Camera Motion Consistency (CamMC). For video v, let $P_{v,t} = [R_{v,t}^{\mathrm{rel}} \mid \mathbf{c}_{v,t}]$ and $\widehat{P}_{v,t} = [\widehat{R}_{v,t}^{\mathrm{rel}} \mid \widehat{\mathbf{c}}_{v,t}]$. Over N videos,

$$\mathrm{CamMC} = \frac{1}{N} \sum_{v=1}^{N} \sum_{t=1}^{T_v} \|P_{v,t} - \widehat{P}_{v,t}\|_F. \tag{43}$$

Table 5: **Independent training runs of the default MeRoPE configuration.** Each entry averages three paired-seed evaluations; Run 1 is used in the main comparison. The final row reports the observed range across training runs.

Training run	rot°↓	tr%↓	AUC@3↑	AUC@10↑	CamMC↓	FID↓	FVD↓
Run 1	0.898	1.662	66.595	87.884	1.415	7.948	82.979
Run 2	0.880	1.612	66.440	87.967	1.380	8.277	83.282
Run 3	0.906	1.668	67.292	88.047	1.423	7.692	79.853
Max–min	0.026	0.056	0.852	0.163	0.043	0.585	3.429

Table 6: **VGGT-Ω on real driving videos.** The 128 original-command clips use the same held-out identities as the pose benchmark. These measurements characterize pose-recovery error, not generated-video performance.

Source	Clips	rot°↓	tr%↓	AUC@3↑	AUC@10↑	CamMC↓
Mixed	128	0.434	1.115	79.284	92.459	0.810

PanShot Rotation Error (RotErr). PanShot sums per-frame geodesic errors in radians, following its official evaluation code (Zhang et al., 2026b):

$$\mathrm{RotErr} = \sum_{t=1}^{T} \arccos\left(\mathrm{clamp}\left(\frac{\mathrm{tr}\left((R_t^{\mathrm{rel}})^\top \widehat{R}_t^{\mathrm{rel}}\right) - 1}{2}, -1, 1\right)\right). \tag{44}$$

Translation Error (TransErr). Used on PanShot with the shared-scale trajectories in Eq. (38):

$$\mathrm{TransErr} = \sum_{t=1}^{T} \|\mathbf{c}_t - \widehat{\mathbf{c}}_t\|_2. \tag{45}$$

C Additional Experimental Analyses

C.1 Training-Seed Variation

Table 5 reports three independent training runs under the same recipe, each averaged over three paired-seed evaluations. The observed CamMC range is 0.043: the main-table gaps to all baselines except RayNova exceed it. FID/FVD differences among MeRoPE, UCPE, and RayNova are within the corresponding training ranges. These ranges provide a reference for interpreting differences.

C.2 Pose-Estimator Reference on Real Videos

Table 6 evaluates VGGT-Ω on real videos using the same pose metrics as for generated videos. These results provide an empirical reference for attainable pose-control performance, accounting for VGGT-Ω's own estimation error.

C.3 Architectural Design and Layer-Wise Modulation

Table 7 varies five design axes around the default MeRoPE configuration. The Comp. rows vary only the adapter compression ratio C. The Design rows fix $C = 4$ and vary one design factor at a time: translation frequency, injection schedule, integration style, or disparity anchor count; check marks denote active mechanisms and blank cells denote inactive ones.

Adapter compression. The Wan2.2 backbone has width 3072 with 24 attention heads, and the ratio C sets the parallel camera-attention branch

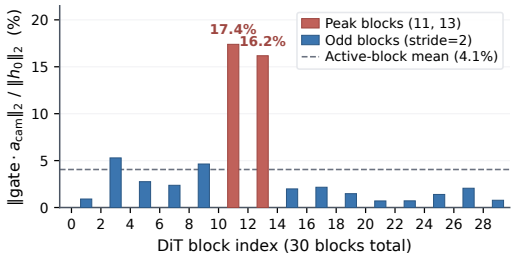

Figure 6: **Per-block camera modulation am-**

Table 7: **Architecture and injection ablations on mixed driving scenes.** Best in **bold**; second-best is underlined within each group.

		Configuration					Overall Metrics				
	Variant	C	Injection	CamSA	ReRoPE	Multi-f	rot°↓	tr%↓	AUC@3↑	AUC@10↑	CamMC↓
Comp.	C=2	2	15 blks	✓		✓	0.900	**1.558**	**66.640**	**87.921**	**1.383**
	C=8	8	15 blks	✓		✓	<u>0.899</u>	<u>1.626</u>	66.219	87.721	<u>1.401</u>
	C=4 (ours)	4	15 blks	✓		✓	**0.898**	1.662	<u>66.595</u>	<u>87.884</u>	1.415
Design	Single f	4	15 blks	✓			0.907	2.663	66.149	87.702	1.832
	Every block	4	30 blks	✓		✓	<u>0.884</u>	<u>1.578</u>	**68.228**	**88.453**	**1.373**
	Peak 11, 13	4	2 blks	✓		✓	**0.873**	1.721	64.975	87.251	<u>1.411</u>
	ReRoPE (Li et al., 2026a)	4	15 blks		✓	✓	1.089	**1.527**	63.983	87.148	1.551
	Two endpoints	4	15 blks	✓		✓	0.925	1.735	61.596	85.997	1.465
	Default (ours)	4	15 blks	✓		✓	0.898	1.662	<u>66.595</u>	<u>87.884</u>	1.415

to width $3072/C$ with $24/C$ heads while preserving the 128-dimensional head size. Thus, $C = 2$, 4, and 8 use widths 1536, 768, and 384, with 12, 6, and 3 heads, respectively. CamMC varies by only 0.032 across these widths, within the observed training-seed range. Increasing adapter capacity therefore does not establish a consistent joint-control gain; we retain $C = 4$ as the default.

Translation frequencies. The Multi-f column indicates whether metric translation uses multiple wavelength bands; the Single-f variant keeps the same 128-dimensional channel allocation and changes only the translation frequency. Using a single frequency increases tr% from 1.662 to 2.663 and worsens CamMC, despite much smaller changes in directional AUC. This supports multiple translation frequencies for improving scale-aligned trajectory agreement, beyond the inclusion of translation PE alone.

Injection schedule. The Injection column compares two selected blocks (Peak 11,13), the default 15 alternating blocks, and all 30 blocks. The nuScenes layer probe in Figure 6 motivates the two-block selection. Using these blocks reduces camera-branch executions by 87% with comparable CamMC, but lowers directional AUC. Conversely, injecting at every block improves AUC while doubling branch executions; its CamMC gain remains within the observed training range.

Integration style. A check in CamSA denotes the dedicated parallel camera self-attention branch; a check in ReRoPE (Li et al., 2026a) denotes carving low-frequency camera rotary channels into the backbone's native self-attention without that branch. ReRoPE improves tr% but worsens rotation accuracy and CamMC, indicating a trade-off rather than a uniform gain. The parallel CamSA branch provides better joint control in this comparison.

Disparity anchor count. Replacing six anchors by the two spherical-arc endpoints retains 36 disparity channels by increasing the triplets per anchor from two to six. Both directional AUC and CamMC worsen, supporting intermediate anchors rather than allocating the same channel budget solely to the endpoints.

C.4 Cost of Relaxing Per-Token Factorization

Strictly per-token encodings transform each token once, whereas MeRoPE expresses metric translation and disparity anchors in each query-camera frame and therefore constructs query-dependent key/value representations. Table 8 measures the resulting cost at both the operator and full-model levels.

Evaluation protocol. The *Single CamSA Block* columns time one complete camera self-attention (CamSA) block, including positional modulation, attention, and output projection, while fixing the

Table 8: **Computational cost of query-camera grouping.** Single-block results use one H20; full-model results use four H20 GPUs with a matched 15-block injection schedule.

		Single CamSA Block		Training		Latent Denoising	
Method	Factorization	Latency (ms)↓	Peak Mem. (GiB)↓	Throughput (clips/s)↑	Peak Mem. (GiB/GPU)↓	Throughput (clips/s)↑	Peak Mem. (GiB/GPU)↓
Wan2.2 base	No CamSA	N/A		5.30	35.80	0.197	29.65
GTA (Miyato et al., 2024)	Per-token	4.97	0.38	4.89	36.58	0.188	31.03
UCPE (Zhang et al., 2026a)	Per-token	5.03	0.40	4.86	36.58	0.188	31.03
MeRoPE	Query-grouped	12.47	1.21	4.70	37.57	0.177	31.13

block dimensions and input tokens across methods. This isolated benchmark uses one H20 because it excludes the backbone and distributed model sharding. All efficiency measurements use 288×512, 49-frame nuScenes inputs. The *Training* and *Latent Denoising* columns instead measure full training steps and the complete denoising loop on four H20 GPUs under the production four-way HSDP setup.

Matched comparison. For the full-model benchmark, GTA, UCPE, and MeRoPE inject CamSA into the same 15 alternating transformer blocks (Blocks $1, 3, \ldots, 29$), yielding 5.14B parameters for each camera-conditioned model; the backbone-only baseline contains 5.00B parameters. This controls both the injection schedule and model size across the three camera encodings.

Results. Query-camera grouping is more expensive within a single CamSA block, but the full backbone amortizes much of this local overhead. Relative to UCPE, MeRoPE requires about $2.5\times$ the isolated-block latency and $3\times$ the peak memory, whereas full-model throughput decreases by 3.3% for training and 5.9% for denoising, with peak-memory increases of 2.7% and 0.3%, respectively. GTA and UCPE have nearly identical system costs under the matched schedule.

Source of overhead and future work. The extra local cost comes from constructing and storing separate transformed K/V tensors for each query-camera group rather than from additional attention dot products. Processing query-camera groups in smaller batches can reduce peak memory, while a fused grouped-attention kernel could avoid expanded K/V tensors and reduce both memory and latency.

C.5 Real-World Grounding with Retrieved History

We conduct history-conditioned experiments on nuScenes (Caesar et al., 2020) and World Engine (Li et al., 2026b), which both provide revisit observations of the same locations. Historical views are retrieved from separate traversals in nuScenes and rendered from 3D Gaussian Splatting scenes in World Engine.

To test grounding beyond the current field of view, we encode calibrated same-location images with the video VAE and inject their patch tokens through two pathways. History CA follows CityRAG (Chou et al., 2026): video tokens attend to retrieved-reference K/V through standard cross-attention. History SA follows Seoul World Model (Seo et al., 2026): retrieved reference tokens are concatenated with the current video tokens and jointly processed by self-attention across successive DiT blocks; only the generated video tokens are decoded.

Training strategy. We use three strategies: (i) randomly drop the clean conditioning frames when usable history is available, encouraging prediction from historical observations; (ii) include off-trajectory historical views to reduce the sampling bias toward high-overlap, nearly aligned views; and (iii) add 100 to the history tokens' temporal position coordinates to distinguish them from the current video. The full conditioning prefix is retained at evaluation.

Table 9 shows that both history-injection pathways improve the reported pose-control metrics, while History SA provides a larger visual-quality gain. Relative to no history, History SA reduces FVD by about 22%, whereas History CA changes it only slightly. Figure 7 complements these metrics: in the two examples, History SA better preserves scene structure visible in the retrieved views at later

Table 9: **Real-world grounding with retrieved history on nuScenes.** Best in **bold**; second-best is underlined.

Method	rot°↓	tr%↓	AUC@3↑	AUC@10↑	CamMC↓	FID↓	FVD↓
MeRoPE (no history)	0.82	2.21	48.76	83.24	1.52	19.82	134.15
MeRoPE + History CA	<u>0.78</u>	<u>2.13</u>	**52.01**	<u>84.17</u>	<u>1.46</u>	<u>19.59</u>	<u>130.27</u>
MeRoPE + History SA	**0.74**	**2.11**	<u>51.65</u>	**84.22**	**1.43**	**18.21**	**104.41**

frames, while History CA remains closer to the no-history result. These observations favor joint self-attention for integrating historical context in this setting.

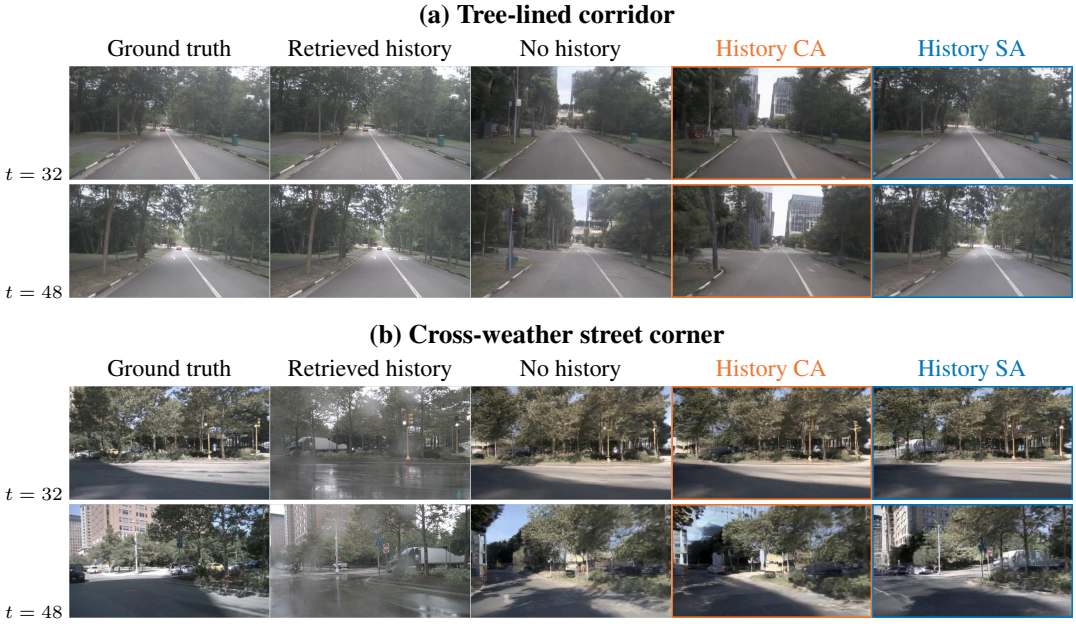

Figure 7: **Qualitative comparison of history injection.** Two nuScenes examples at $t = 32$ and $t = 48$. Across the same scenes and timesteps, History SA preserves retrieval-supported late-frame geometry, whereas History CA follows the no-history layout.

D Additional Qualitative Results

D.1 PanShot Across Camera Optics



Figure 8: **PanShot generations across camera optics.** Rows 1–4 span (FoV, ξ) from $(97°, 0.00)$ to $(173°, 1.66)$, covering pinhole through fisheye cameras.

D.2 Matched-Backbone Comparison with Prior Camera Encodings

Figure 9 presents matched-backbone examples from nuScenes, AV2, and Waymo, retained from the earlier training protocol. In these clips, MeRoPE preserves more coherent late-frame geometry.

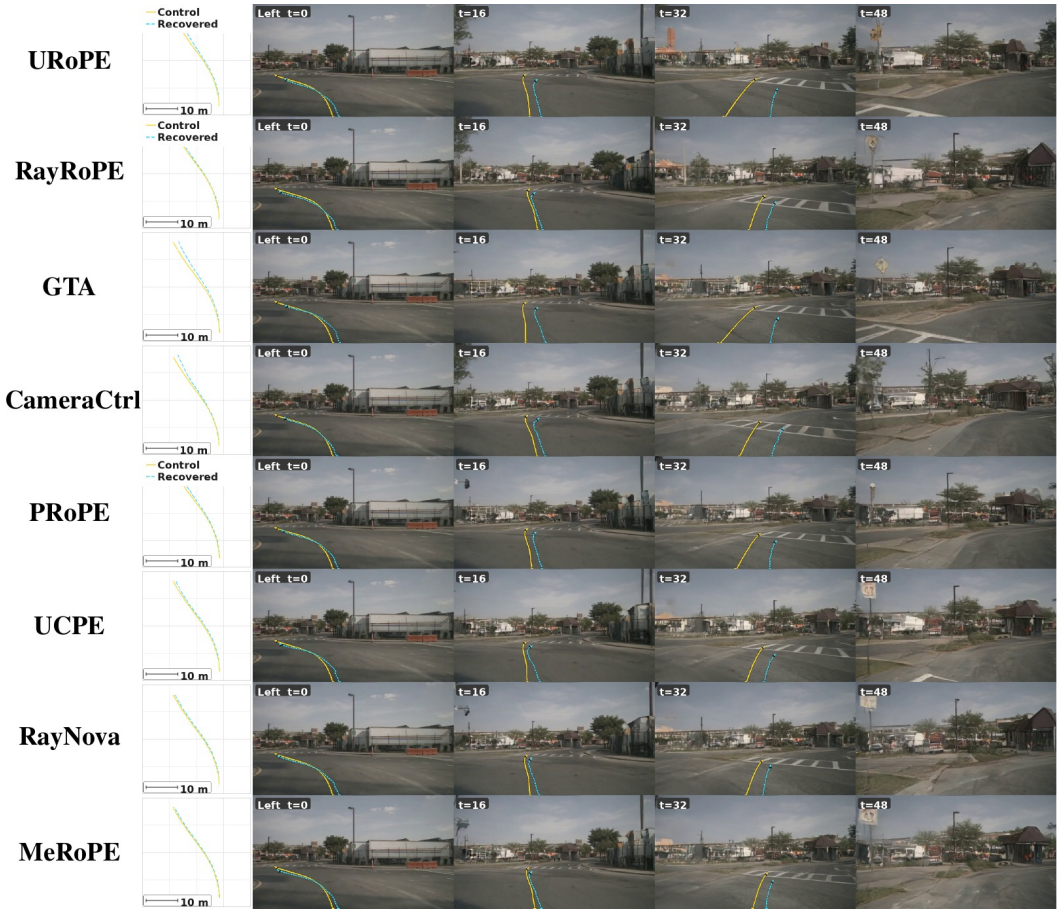

Figure 9: **Matched-backbone qualitative comparison on nuScenes (1/3).** Each row shows commanded and VGGT-Ω-recovered paths with frames at $t \in \{0, 16, 32, 48\}$.

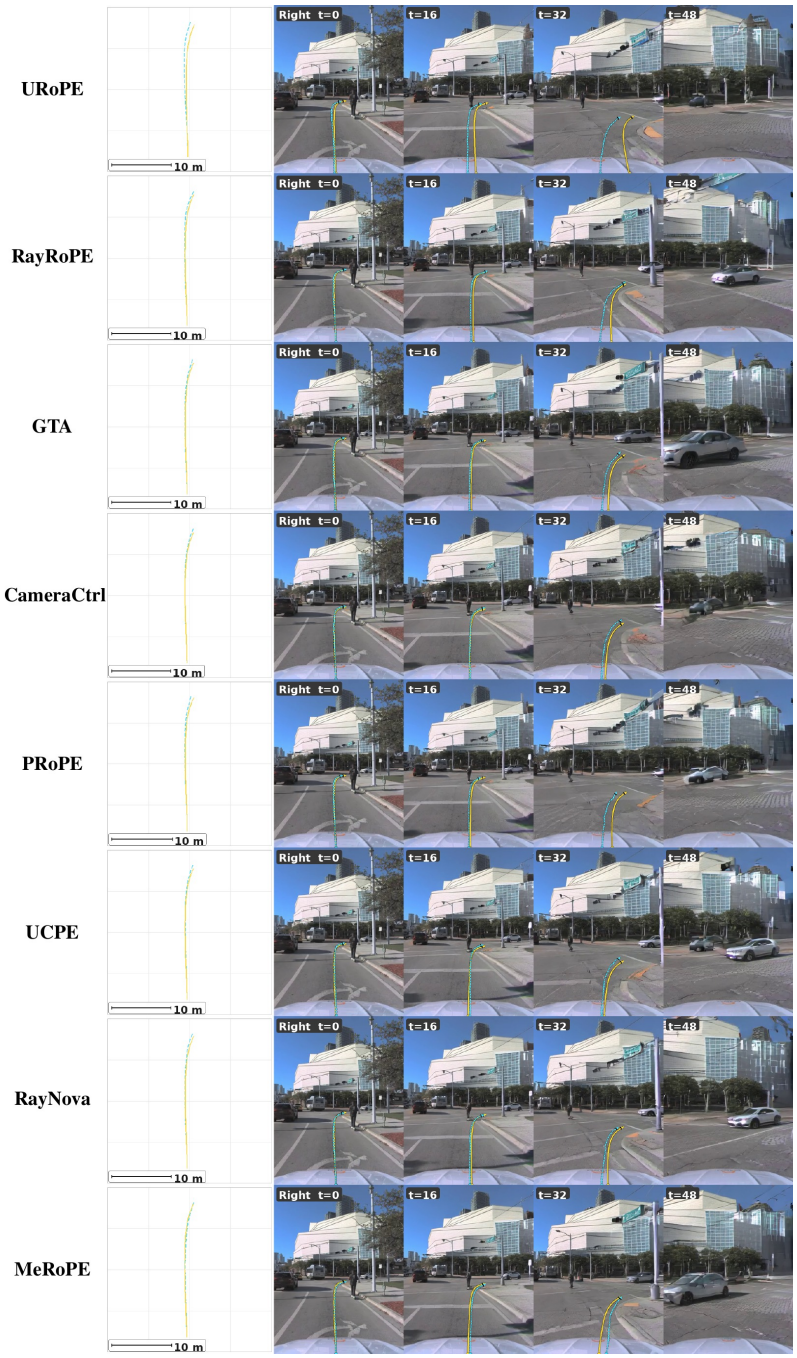

Figure 9: **Matched-backbone qualitative comparison on AV2 (2/3).** The same protocol is shown for an AV2 scene.

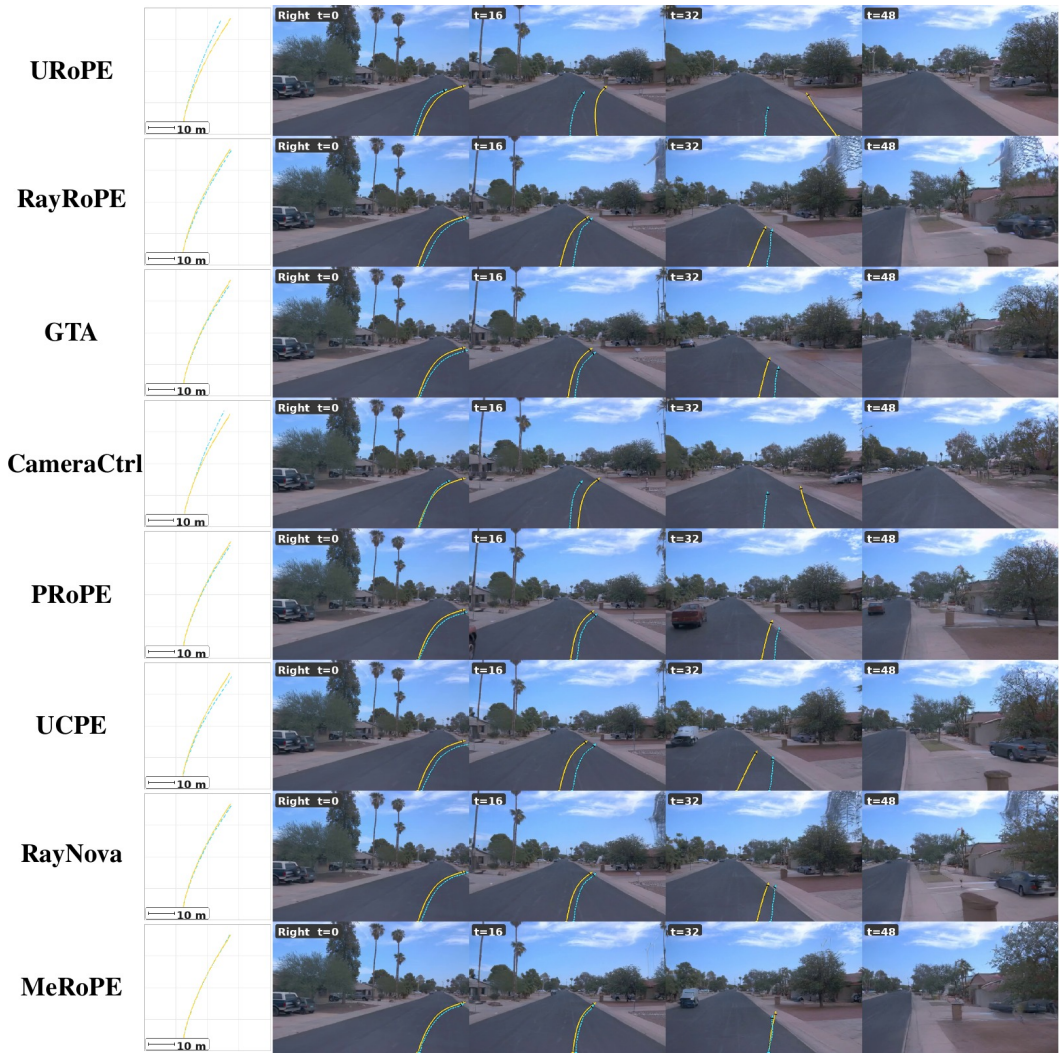

Figure 9: **Matched-backbone qualitative comparison on Waymo (3/3).** The same protocol is shown for a Waymo scene.