BEYOND THE CENTRAL RAY: RAY DIFFERENTIAL ATTENTION FOR NOVEL VIEW SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalizable novel-view synthesis reconstructs unseen views from posed images using transformer in a single forward pass. Transformers tackle this task by aggregating features across views, where performance critically depends on how attention encodes camera geometry. Existing methods predominantly represent each token as an infinitesimal central ray, discarding the patch's local angular scale, anisotropy, and shear across the viewing sphere. To address this limitation, we introduce **Ray Differential Attention (RDA)**. First, we represent each token by a ray differential—a unit ray along with its spatial image derivatives capturing the local patch footprint without requiring scene depth. Second, RDA incorporates this geometry into both attention routing and feature aggregation. For routing, dual-primal query–key transforms yield relative, world-frame-invariant attention scores. For aggregation, an invertible value transform enables consistent cross-view content transport via an exact query-side inverse. This establishes a broader principle that camera geometry should actively govern both token interaction and feature expression, rather than serving as passive conditioning. In a controlled LVSM host, RDA outperforms five camera baselines across PSNR, SSIM, and LPIPS on RealEstate10K and DL3DV. Rigorous ablations confirm that both the full ray differential footprint and the dual read/write pathways are essential to these gains.

1 INTRODUCTION

Novel-view synthesis reconstructs what a scene would look like from an unseen camera given a small set of posed images. It supports free-viewpoint rendering, immersive content creation, and visual understanding for embodied agents. In the generalizable setting, one model must render a new scene in a single forward pass, without optimizing a scene-specific representation at test time. Recent Transformers make this setting especially attractive because they can gather evidence across views with little built-in 3D structure (Sajjadi et al., 2022; Varma T et al., 2023; Jin et al., 2024). Their performance, however, depends on how attention understands the camera geometry of each image token.

There is no established solution to this geometric interface. Relative pose describes how cameras are arranged, but pose alone does not specify how pixels map to viewing directions. Projective and ray-based encodings add intrinsics or token-level sightlines (Kong et al., 2024; Miyato et al., 2024; Li et al., 2025; Zhang et al., 2026a; Xie et al., 2026), yet camera geometry still cannot determine which two tokens observe the same surface point because scene depth is unknown. A central ray improves on pose-only encodings but reduces a finite patch to one direction. Because its local scale, anisotropy, and shear depend on intrinsics, equal poses and principal rays need not imply equal footprints. An orthonormal completion retains orientation but discards these metric differences, leaving attention to infer local calibration from appearance.

Our key observation is that the geometry of an image token is described more completely by a ray together with how that ray changes across the image. The horizontal and vertical derivatives of a unit ray record the angular size, anisotropy, and shear of the token's local footprint. Together, the ray and its two derivatives form a ray differential, a classical tool for footprint estimation in rendering (Igehy, 1999). Attention first compares tokens to obtain weights and then combines their features; both operations should use camera geometry consistently, and neither should depend on an arbitrary choice

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.

of world coordinates. Simply appending a ray differential to the token provides no such guarantee and leaves the network to discover these operations from data. Figure 1 summarizes this motivation and our solution.

We introduce **Ray Differential Attention (RDA)** to turn this observation into a complete attention layer. For each token, RDA computes the calibrated unit ray and its horizontal and vertical derivatives, rotates them with the camera pose, and includes the camera center in an invertible transform. When computing attention scores, it applies the inverse transpose of this transform to the query and the direct transform to the key. Their dot product therefore depends on the relative geometry between the two tokens and remains unchanged when the shared world frame is relabeled. When combining content, RDA transforms each value before aggregation and applies the exact inverse associated with the query afterward. This construction keeps the original attention width and requires neither ground-truth nor predicted depth.

The method works by assigning known camera geometry and unknown scene geometry different roles. The ray differential supplies the local calibration that follows directly from the camera, while image features remain responsible for inferring correspondence and depth. Relative query–key transforms use that calibration to decide where to read, and the invertible value transform ensures that content read from different cameras is expressed consistently at the query token.

Experiments use the same LVSM host to compare RDA with five depth-free camera methods on RealEstate10K and DL3DV under matched training and evaluation protocols. Controlled studies support both the full ray differential and its use in the score and content parts of attention. Additional crop-and-resize tests and same-device profiling examine behavior under calibration changes and the computational cost of the complete method.

Our contributions are threefold:

- We identify the local ray footprint as a missing piece of token-level camera information in central-ray encodings and represent it with a depth-free ray differential that retains angular scale, anisotropy, and shear.
- We introduce RDA, which turns this representation into relative, world-frame-invariant attention scores and consistently combines values through an exact inverse, without increasing the score width (Section 3).
- We establish a geometry-aware attention principle that camera information should guide both token interaction (reading) and feature transformation (writing), rather than only serving as positional conditioning.

2 Related Work

2.1 Feed-forward view synthesis

Generalizable novel-view synthesis has evolved from geometry-structured aggregation (Wang et al., 2021; Varma T et al., 2023) and explicit 3D reconstruction (Li et al., 2024; Zhang et al., 2024) toward token-based rendering with minimal 3D inductive biases. In the latter paradigm, SRT compresses multi-view inputs into a latent set, while LVSM employs a decoder-only Transformer to map posed image tokens directly to target views (Sajjadi et al., 2022; Jin et al., 2024). Recent variants relax camera assumptions or reorganize interaction topologies, such as estimating cameras from unposed inputs (Jiang et al., 2025) or decoupling cross-view attention (Jia et al., 2026). Because these architectures alter the supervision, renderer, and scene representation concurrently, they do not isolate the role of camera encoding. We instead keep LVSM's renderer and attention topology fixed to study that geometric interface directly.

2.2 Relative geometry in attention

Relative-position attention ensures token interactions depend on coordinate differences rather than arbitrary global frames (Shaw et al., 2018; Su et al., 2024; Zhang et al., 2026b). In novel-view synthesis, early formulations incorporated geometric relations via ray distance biases or relative poses (Suhail et al., 2022; Venkat et al., 2023; Safin et al., 2023). Recent camera-aware operators formalize

this at the pose, projective, or channel level. At the pose level, CaPE conditions attention on relative motion (Kong et al., 2024), whereas GTA applies group representations across queries, keys, and values to map aggregated features back to the query frame (Miyato et al., 2024). Projective methods like PRoPE and From Rays to Projections incorporate camera matrices or target-view projections (Li et al., 2025; Wu et al., 2026b). Concurrently, ReRoPE and DPPE highlight the importance of channel allocation, showing that separating rotation and translation prevents training interference (Li et al., 2026; Kenney & Suzuki, 2026). Together, this literature separates three design choices: the camera object, the relative action, and the channel split. RDA uses distinct score and content actions, but its primary departure lies in the first choice: it models the calibrated *neighborhood* of a ray rather than only its center.

2.3 Ray-space token geometry

Ray-based methods diverge primarily in whether they model an isolated sightline or introduce structure along it. A central ray is commonly parametrized by its camera origin and direction via Plücker coordinates (Jin et al., 2024); UCPE augments this direction with an orthonormal frame, while ViewRope and SCoPE inject per-patch directions alongside image-space RoPE (Zhang et al., 2026a; Xiang et al., 2026; Yin et al., 2026). Alternatively, other approaches introduce locations along the sightline: URoPE projects fixed depth anchors into the query view, and RayRoPE predicts a point with learned uncertainty (Xie et al., 2026; Wu et al., 2026a). Despite their differences, these representations treat patches as infinitesimal directions or discrete 3D points, discarding how neighboring sightlines vary across the token. Ray differentials (Igehy, 1999) supply this missing first-order variation, preserving the local angular scale, anisotropy, and shear induced by intrinsics. RDA brings this classical tool into multi-view attention, utilizing it both to route attention scores and to transport features between token frames without selecting or predicting scene depth.

3 Ray Differential Attention

3.1 Problem setup and method overview

Preliminaries. Given M context images I_m with intrinsics K_m and camera-to-world poses $(R_m^{\mathrm{cam}}, t_m)$, together with a target camera $(K_\star, R_\star^{\mathrm{cam}}, t_\star)$, the task is to predict the target RGB image $\widehat{I}_\star$. We consider generalizable, depth-free novel-view synthesis: a shared model renders $\widehat{I}_\star$ in one forward pass, without scene-specific optimization, ground-truth depth, or predicted depth. The LVSM host processes tokens associated with the posed context and target views. At each self-attention layer, token i provides $q_i, k_i, v_i \in \mathbb{R}^{144}$ and a camera packet containing a camera-to-world rotation $R_i^{\mathrm{cam}} \in \mathrm{SO}(3)$, camera center $t_i \in \mathbb{R}^3$, intrinsics $K_i \in \mathbb{R}^{3\times3}$, pixel center $u_i = (x_i, y_i)$, and patch-grid coordinate $p_i = (p_{i,x}, p_{i,y})$. Here $\mathrm{SO}(3)$ denotes the group of three-dimensional rotation matrices. RDA changes only how this known camera packet acts inside attention; correspondence and scene geometry remain encoded by the learned image features.

Notation. All vectors are column vectors. We write I_n for the $n \times n$ identity matrix and omit the subscript when its size is clear. Superscripts T and $-T$ denote transpose and inverse transpose, respectively, and $\|\cdot\|$ denotes the Euclidean norm. The symbols $\otimes$ and $\oplus$ denote the Kronecker product and block-diagonal direct sum, respectively.

Motivation. Existing ray-based attention leaves two gaps: a central ray omits finite-patch calibration, and Q/K-only geometry does not define how retrieved values should be combined across cameras. RDA addresses both with a calibrated ray differential used for score routing and content transport.

Overview. Figure 1 shows how the calibrated token transform enters both parts of attention. The score path applies dual and primal query/key actions to expose pairwise relative geometry (Section 3.3). The content path transports source values and maps their aggregate into the query representation (Section 3.4). Both preserve standard attention and the host channel width (Section 3.5).

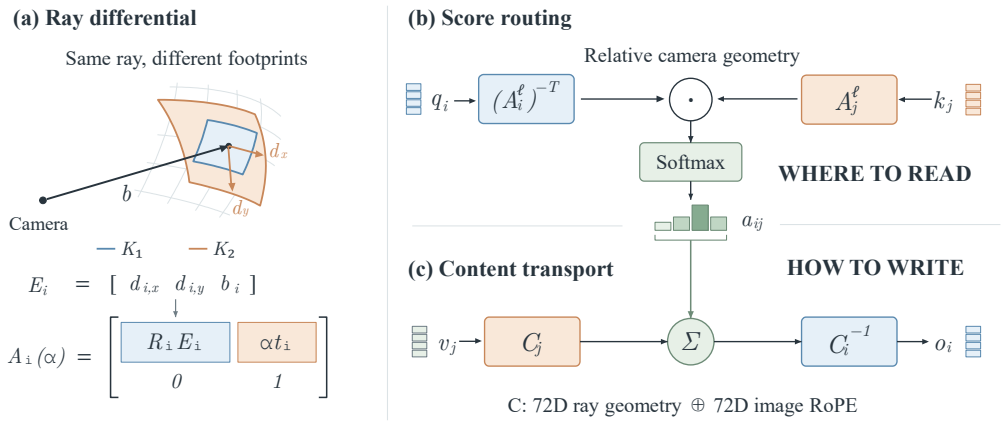


Figure 1: Overview of RDA. A calibrated ray differential distinguishes token footprints that share a central ray. Combined with camera pose, it defines $A_i(\alpha)$ for score routing and contributes to the content carrier C_i for query-local transport; α controls translation scale. Sections 3.2–3.4 specify both paths.

3.2 From an image patch to a calibrated ray differential

An image token covers a finite patch whose neighboring directions vary with focal length, aspect ratio, skew, and image location. We therefore pair its central ray with the first-order image variation that captures this local footprint. Figure 2 compares the information retained by a unit ray, an orthonormal ray frame, and the full ray differential.

Let $\bar{u}_i = [x_i, y_i, 1]^T$ be the homogeneous pixel center. We first back-project and normalize it to obtain the camera-local bearing

$$r_i = K_i^{-1}\bar{u}_i, \qquad b_i = \frac{r_i}{\|r_i\|}. \tag{1}$$

To recover the patch footprint around this bearing, we differentiate b_i with respect to image coordinates normalized to $[-1, 1]^2$. With $e_x = (1, 0, 0)^T$, $e_y = (0, 1, 0)^T$, image width W, and image height H, the two tangent directions are

$$d_{i,x} = \frac{W}{2\|r_i\|}(I_3 - b_i b_i^T)K_i^{-1}e_x, \qquad d_{i,y} = \frac{H}{2\|r_i\|}(I_3 - b_i b_i^T)K_i^{-1}e_y. \tag{2}$$

The projector $I_3 - b_i b_i^T$ removes the component parallel to the unit ray, so both derivatives lie in the tangent plane of the viewing sphere. We collect this first-order information as

$$E_i = [d_{i,x}, d_{i,y}, b_i] \in \mathbb{R}^{3\times 3}. \tag{3}$$

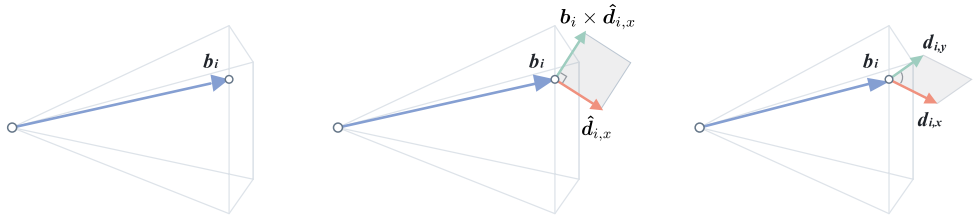
At regular pinhole-camera locations, the two derivatives span the tangent plane and b_i supplies its normal, making E_i nonsingular.

The non-orthonormal structure of E_i is deliberate. A unit ray retains only direction. An orthonormal completion additionally uses $\widehat{d}_{i,x} = d_{i,x}/\|d_{i,x}\|$ and the three-dimensional cross product $b_i \times \widehat{d}_{i,x}$ to retain an orientation around the ray. Both representations discard the derivative magnitudes and their correlation. In contrast, the first two columns of E_i preserve the local angular scale, anisotropy, and shear induced by calibration. This information is available before observing the scene and does not introduce a depth or a hypothesized correspondence.

Finally, we orient the differential in the shared world frame and attach the camera center:

$$L_i = R_i^{\text{cam}} E_i, \qquad A_i(\alpha) = \begin{bmatrix} L_i & \alpha t_i \\ 0 & 1 \end{bmatrix}. \tag{4}$$

This homogeneous matrix is the interface between camera geometry and the two attention paths. The score path uses one learned translation scale α_ℓ per layer, initialized to 0.01, so different layers can tune the influence of camera displacement relative to viewing geometry. The content path uses the fully geometric transform $A_i = A_i(1)$, with no learned camera parameter.



$b_i = r_i / ||r_i||$ **(a) Unit ray**

$F_i = \left[\hat{d}_{i,x},\ b_i \times \hat{d}_{i,x},\ b_i\right] \in SO(3)$ **(b) Orthonormal ray frame**

$E_i = [d_{i,x},\ d_{i,y},\ b_i] \in \mathbb{R}^{3\times 3}$ **(c) Full ray differential (RDA)**

Figure 2: Comparison of token-local ray representations. A unit ray retains only viewing direction. An orthonormal completion additionally retains orientation but normalizes away the local metric. The full ray differential used by RDA also preserves angular scale, anisotropy, and cross-axis correlation.

3.3 Where to read: relative score routing

Attention should respond to the geometry between a query and a key, rather than to the arbitrary world frame in which both cameras happen to be reported. Feeding absolute camera parameters to the token does not enforce this behavior. RDA instead makes relativity a property of the dot product by pairing an inverse-transpose action on the query with a direct action on the key.

For layer ℓ, denote $A_i^\ell = A_i(\alpha_\ell)$ and reshape each 144D query and key into 36 four-dimensional blocks. The matrix $I_{36} \otimes A_i^\ell$ is block diagonal with 36 copies of A_i^ℓ, so it applies the same token transform independently to every block. We use the inverse transpose for queries and the direct transform for keys:

$$\widehat{q}_i = \left(I_{36} \otimes (A_i^\ell)^{-T}\right) q_i, \qquad \widehat{k}_i = \left(I_{36} \otimes A_i^\ell\right) k_i. \tag{5}$$

The transformed features are passed to the standard dot-product attention. For query i and key j, this gives

$$s_{ij} = \frac{\widehat{q}_i^T \widehat{k}_j}{\sqrt{144}} = \frac{q_i^T \left(I_{36} \otimes (A_i^\ell)^{-1} A_j^\ell\right) k_j}{\sqrt{144}}. \tag{6}$$

Thus the image features still decide whether two tokens match, while the camera-dependent part of their comparison is exposed only through the relative transform $(A_i^\ell)^{-1} A_j^\ell$.

This factorization gives two useful properties. First, a common rigid relabeling of the world frame left-multiplies every A_i^ℓ by the same matrix, which cancels from the relative product. The logits and attention weights are therefore unchanged. Second, the implementation transforms each token once before the usual matrix multiplication; it does not need to materialize a pairwise tensor of relative camera matrices. Repeating a 4×4 action over existing channel blocks also preserves the 144D score width. Appendix A states the assumptions and proves the invariance result.

3.4 How to write: query-local content transport

Relative scores settle which source tokens contribute, but they do not specify how source features should be combined. If camera geometry acts only on queries and keys, the weighted value sum has no corresponding transformation from the source camera to the query camera. RDA supplies this missing step with a square transport: each source prepares its value for aggregation, and the query maps the aggregate back to its own representation with a known inverse.

The content carrier keeps two complementary coordinate systems within the 144 value channels. The first 72 channels form 18 four-dimensional affine blocks. The remaining 72 channels split evenly between horizontal and vertical image position. Let $\mathcal{R}_x^{36}(p_{i,x})$ and $\mathcal{R}_y^{36}(p_{i,y})$ be the corresponding 36×36 rotary matrices. With $A_i = A_i(1)$, the term $I_{18} \otimes A_i^{-T}$ applies the affine inverse transpose to the first 18 blocks, while $\oplus$ assembles this 72×72 geometry map and the two rotary maps into one block-diagonal 144×144 carrier:

$$C_i = \left(I_{18} \otimes A_i^{-T}\right) \oplus \mathcal{R}_x^{36}(p_{i,x}) \oplus \mathcal{R}_y^{36}(p_{i,y}). \tag{7}$$

The source-side operation $\widehat{v}_j = C_j v_j$ places the value in the transport carrier. Because the carrier is a direct sum of invertible blocks, the query-side map is available analytically:

$$C_i^{-1} = \left(I_{18} \otimes A_i^{T}\right) \oplus \mathcal{R}_x^{36}(p_{i,x})^{-1} \oplus \mathcal{R}_y^{36}(p_{i,y})^{-1}. \tag{8}$$

The resulting contribution from source j to query i is expressed through $C_i^{-1} C_j v_j$. This relative form explains why the transport is stable under coordinate changes: under a common world-frame relabeling, all source carriers acquire the same left factor on their geometry channels and the query inverse cancels it. It also provides an exact local round trip, $C_i^{-1} C_i = I_{144}$ before floating-point error, rather than asking a learned decoder to undo the source transform. The split carrier preserves the host value width while retaining image-grid position alongside calibrated multi-view geometry.

3.5 Complete RDA attention

The full layer now follows the pipeline in Figure 1. Camera preprocessing produces E_i and L_i for every token; these quantities can be reused across layers. At layer ℓ, A_i^ℓ produces $\widehat{q}_i$ and $\widehat{k}_i$, while the fixed content carrier produces $\widehat{v}_i = C_i v_i$. Standard softmax attention computes $a_{ij} = \mathrm{softmax}_j(s_{ij})$, after which the query-side inverse returns the weighted content to token i:

$$y_i = C_i^{-1} \sum_j a_{ij} C_j v_j = C_i^{-1} \sum_j \mathrm{softmax}_j\left(\frac{\widehat{q}_i^{T} \widehat{k}_j}{\sqrt{144}}\right) C_j v_j. \tag{9}$$

The score path therefore controls *where* token i reads, and the content path controls *how* what it reads is written into token i.

Together, the relative score identity and the relative content transport make the camera-dependent part of the complete output unchanged under a common rigid relabeling of all cameras, assuming E_i is nonsingular and up to floating-point error. This guarantee does not prescribe scene correspondence: appearance features remain responsible for resolving which rays meet at the same unknown surface. Appendix A gives the complete proof, and Appendix A.5 describes the wider score constructions used only as ablations.

4 Experiments

Our experiments address three questions. First, does RDA improve over existing depth-free camera operators under a controlled rendering host? Second, which parts of RDA account for the improvement? Third, do the gains persist on a second dataset and under explicit changes to image sampling and camera intrinsics? We answer these questions with two three-seed benchmark comparisons, matched pathway and representation ablations, qualitative reconstructions, a paired crop/resize stress test, and same-device profiling.

4.1 Experimental setup

Datasets and metrics. RealEstate10K is our primary benchmark. We evaluate the predeclared 20,000-update checkpoint on a fixed 5,899-scene panel using two context views, three target views, and 256×256 images. To test the method in a second data regime, we evaluate completed 100,000-update checkpoints on a frozen 140-scene DL3DV panel with two context views, ten target views, and a 100-frame span at the same resolution. Scenes receive equal weight. We report PSNR and SSIM, where higher is better, and LPIPS, where lower is better.

Table 1: RealEstate10K results on the fixed 5,899-scene panel at 20,000 updates. Entries are mean ± sample standard deviation over seeds 0, 1, and 2; the best mean in each column is shaded light green.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
CaPE	20.643 ± 0.158	0.6333 ± 0.0077	0.2673 ± 0.0073
PRoPE	21.343 ± 0.066	0.6693 ± 0.0025	0.2341 ± 0.0105
GTA	21.185 ± 0.056	0.6564 ± 0.0018	0.2488 ± 0.0031
URoPE	20.226 ± 0.055	0.6272 ± 0.0030	0.2706 ± 0.0087
UCPE	21.436 ± 0.187	0.6733 ± 0.0023	0.2244 ± 0.0017
RDA (ours)	21.882 ± 0.129	0.6847 ± 0.0039	0.2162 ± 0.0010

Training and evaluation protocol. Every method uses the same six-layer, eight-head LVSM host, 144D learned head width, and FP16 training. Within each seed, methods share the host initialization and scene order. We evaluate seeds 0, 1, and 2 at the fixed endpoint specified for each dataset, aggregate the three seed-level test means, and report their mean and sample standard deviation. Test performance is never used for checkpoint or hyperparameter selection, and every evaluated checkpoint must pass strict parameter loading on its frozen panel.

Compared methods. We compare RDA with CaPE, PRoPE, GTA, URoPE, and UCPE, covering pose-level, projective, and ray-level camera operators. The rendering backbone, data, optimization, and evaluation are held fixed while the camera operator changes; each baseline retains its published channel allocation in the common host. The comparison is restricted to depth-free methods. Methods that predict scene geometry or depth use additional information and define a different experimental setting.

4.2 Comparison with existing camera operators

RealEstate10K. RDA achieves the best three-seed mean on all three reconstruction metrics (Table 1). UCPE is the strongest external baseline in each column. Relative to UCPE, RDA improves PSNR by 0.446 dB and SSIM by 0.0113, while reducing LPIPS by 0.0083. The improvement therefore appears in both distortion-oriented and perceptual measures rather than trading one for the other. Per-seed results are provided in Appendix C.

Figure 3 complements the aggregate metrics with four held-out targets. Across these examples, RDA better preserves object boundaries and local image structure in regions where several alternatives produce visibly smoother reconstructions. These examples are qualitative diagnostics; the fixed-panel metrics in Table 1 provide the aggregate comparison.

DL3DV. We next test the same operators on a second dataset using ten target views over a 100-frame span. RDA again obtains the best mean PSNR, SSIM, and LPIPS (Table 2). Against the strongest external value in each column, it improves PSNR by 0.011 dB over UCPE, improves SSIM by 0.0023 over UCPE, and reduces LPIPS by 0.0170 relative to GTA. The smaller PSNR margin than on RealEstate10K cautions against assuming a fixed gain across datasets, while the consistent direction across all three metrics supports the method beyond the primary benchmark. Appendix D reports the individual seeds.

4.3 Ablation studies

Score routing and content transport. To separate the two roles of geometry in RDA, we run a matched 2×2 factorial that independently enables the score and content paths. All four seed-0 runs use the CamRay token input, the same initialization and training schedule, and 144D Q/K/V/O widths. A disabled path uses an explicit identity on the same channel allocation, so neither upstream camera input nor model capacity changes. The RDA/RDA cell reuses the seed-0 endpoint from the main comparison; the other three cells are freshly trained controls.

Enabling either path substantially improved all three metrics relative to the identity/identity host (Table 3). The content-only model was closer to full RDA than the score-only model, suggesting that query-local value transport accounts for most of the gain. Adding score routing still improved all

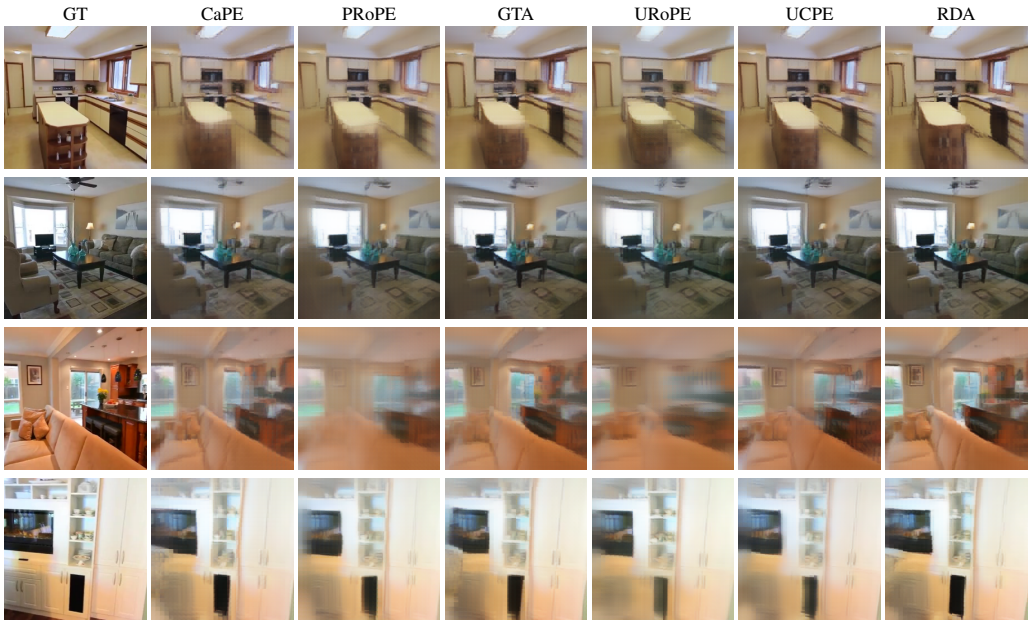


Figure 3: Four representative RealEstate10K test targets. Every row uses the same two context views and target camera for all methods; columns retain a fixed method order for direct visual comparison.

Table 2: DL3DV results on the fixed 140-scene panel at 100,000 updates, using two context views and ten targets over a 100-frame span. Entries are mean ± sample standard deviation over seeds 0, 1, and 2; the best mean in each column is shaded light green.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
CaPE	14.492 ± 0.101	0.3500 ± 0.0044	0.7752 ± 0.0329
PRoPE	14.636 ± 0.007	0.3580 ± 0.0004	0.7325 ± 0.0042
GTA	14.637 ± 0.022	0.3569 ± 0.0009	0.7291 ± 0.0055
URoPE	14.474 ± 0.034	0.3537 ± 0.0013	0.7772 ± 0.0094
UCPE	14.702 ± 0.009	0.3585 ± 0.0001	0.7375 ± 0.0022
RDA (ours)	14.713 ± 0.005	0.3608 ± 0.0001	0.7121 ± 0.0048

three metrics. Because the combined gain was strongly sub-additive and each cell was a separately trained single-seed run, we treat this result as directional evidence that the paths carry overlapping geometric information, not as a precise causal decomposition.

Ray representation. Table 4 changes only the ray object used by the matched score and content transforms. Replacing a unit ray with an orthonormal frame improves all three metrics, showing that orientation around the ray matters. Retaining the full non-orthonormal differential improves further: relative to the orthonormal frame, it raises PSNR by 0.115 dB and reduces LPIPS by 0.0088, with a simultaneous SSIM increase. The result supports the specific information RDA is designed to retain: local angular scale, anisotropy, and shear. Appendix A.5 reports the wider spectral score controls.

4.4 Calibration stress test under crop and resize

The motivation for ray differentials is most directly probed when image sampling and intrinsics change together. On a fixed 128-scene panel, we apply centered crops at scales 0.875 and 0.75 and one-sided left, right, top, and bottom crops at scale 0.875. Each crop is resized to 256×256, and focal length and principal point are updated using the same pixel-center convention. For every scene and method, we pair the cropped result with its canonical evaluation before averaging over the six interventions.

Table 3: Seed-0 score/content factorial on the fixed RealEstate10K panel at 20,000 updates. "Host" and "RDA" denote identity and geometry-aware actions, respectively; all Q/K/V/O widths are 144, and the best result in each metric is shaded light green.

Score	Content	Q/K	V/O	PSNR ↑	SSIM ↑	LPIPS ↓
Host	Host	144	144	16.677	0.5213	0.5482
RDA	Host	144	144	21.136	0.6607	0.2406
Host	RDA	144	144	21.655	0.6793	0.2229
RDA	RDA	144	144	21.734	0.6802	0.2159

Table 4: Matched seed-0 ray-representation ablation on RealEstate10K at 20,000 updates. The score and content transforms change only in the ray object they use; the best metric in each column is shaded.

Configuration	Score dim.	PSNR ↑	SSIM ↑	LPIPS ↓
Unit ray	144	21.558	0.6761	0.2301
Orthonormal frame	144	21.619	0.6784	0.2247
Full differential	144	21.734	0.6802	0.2159

RDA remains the strongest method on all three crop-averaged metrics (Table 5). Across the compared methods, RDA obtains the highest PSNR and SSIM and the lowest LPIPS. All methods improve on their canonical PSNR and SSIM under these crops, so the sign of Δ alone should not be interpreted as robustness. The meaningful comparison is the ranking under the same paired intervention. Moreover, cropping changes RGB sampling and intrinsics jointly, so this experiment tests their coupled effect rather than isolating calibration.

Table 5: Paired crop/resize test on a fixed 128-scene RealEstate10K panel, averaged equally over six interventions. Δ is crop minus canonical; positive PSNR/SSIM and negative LPIPS indicate improvement. The best absolute metric is shaded light green.

Method	Crop PSNR ↑	ΔPSNR	Crop SSIM ↑	ΔSSIM	Crop LPIPS ↓	ΔLPIPS
CaPE	21.154	+0.617	0.66180	+0.02894	0.25378	-0.01770
PRoPE	21.661	+0.716	0.69248	+0.03240	0.22518	-0.02890
GTA	21.397	+0.556	0.67316	+0.02786	0.23743	-0.02613
URoPE	20.331	+0.384	0.64799	+0.02571	0.26300	-0.02077
UCPE	21.480	+0.593	0.69045	+0.02760	0.20974	-0.02421
RDA	22.223	+0.840	0.70506	+0.03382	0.20300	-0.02386

5 Discussion and Conclusion

We introduced RDA to expose the local footprint of an image token to depth-free multi-view attention. Ray differentials retain angular scale, anisotropy, and shear; paired query–key transforms form relative, world-frame-invariant scores, while a query-side inverse transports aggregated content back to the query. In a controlled LVSM host, RDA achieves the best three-seed means on all metrics across both datasets and retains its PSNR/LPIPS advantage under crop/resize. Ablations favor the full differential, show that both paths contribute, and find no benefit from a wider spectral score. Together, these results support using camera geometry to govern both token comparison and content transport.

The study is limited to predominantly static scenes captured by calibrated perspective cameras, with two context views, 256×256 rendering, and one Transformer host. Its conclusions do not yet cover dynamic or deformable scenes, unknown or noisy calibration, non-pinhole optics, substantially higher resolutions, larger context sets, or other rendering backbones. Future work should extend RDA to these settings.

AI USE STATEMENT

Generative AI tools, specifically large language models, were used solely for language polishing and editorial assistance, including improving grammar, clarity, concision, and organization. All AI-assisted revisions were reviewed and, where necessary, rewritten by the authors. The authors take full responsibility for the paper's technical content, claims, citations, experimental results, and conclusions.

REFERENCES

Homan Igehy. Tracing ray differentials. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 179–186, 1999. doi: 10.1145/311535.311555. URL https://doi.org/10.1145/311535.311555.

Xiaosong Jia, Yihang Sun, Junqi You, Songbur Wong, Zichen Zou, Junchi Yan, Zuxuan Wu, and Yu-Gang Jiang. Efficient-LVSM: Faster, cheaper, and better large view synthesis model via decoupled co-refinement attention. *arXiv preprint arXiv:2602.06478*, 2026. URL https://arxiv.org/abs/2602.06478.

Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, and Georgios Pavlakos. RayZer: A self-supervised large view synthesis model. In *IEEE/CVF International Conference on Computer Vision*, pp. 4918–4929, 2025. URL https://arxiv.org/abs/2505.00702.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. LVSM: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242*, 2024. URL https://arxiv.org/abs/2410.17242.

Shun Kenney and Teppei Suzuki. DPPE: Rethinking camera-based positional encoding for scaling multi-view transformers. *arXiv preprint arXiv:2606.31585*, 2026. URL https://arxiv.org/abs/2606.31585.

Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J. Davison. EscherNet: A generative model for scalable view synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 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. *arXiv preprint arXiv:2602.08068*, 2026. URL https://arxiv.org/abs/2602.08068.

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3D: Fast text-to-3d with sparse-view generation and large reconstruction model. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.06214.

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*, 2025. URL https://arxiv.org/abs/2507.10496.

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.

Aleksandr Safin, Daniel Duckworth, and Mehdi S. M. Sajjadi. RePAST: Relative pose attention scene representation transformer. *arXiv preprint arXiv:2304.00947*, 2023. URL https://arxiv.org/abs/2304.00947.

Mehdi S. M. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lucic, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas Funkhouser, and Andrea Tagliasacchi. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2111.13152.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2018. URL https://arxiv.org/abs/1803.02155.

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

Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2207.10662.

Mukund Varma T, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Is attention all that NeRF needs? In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2207.13298.

Naveen Venkat, Mayank Agarwal, Maneesh Singh, and Shubham Tulsiani. Geometry-biased transformers for novel view synthesis. *arXiv preprint arXiv:2301.04650*, 2023. URL https://arxiv.org/abs/2301.04650.

Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. IBRNet: Learning multi-view image-based rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2102.13090.

Yu Wu, Minsik Jeon, Jen-Hao Rick Chang, Oncel Tuzel, and Shubham Tulsiani. RayRoPE: Projective ray positional encoding for multi-view attention. *arXiv preprint arXiv:2601.15275*, 2026a. URL https://arxiv.org/abs/2601.15275.

Zirui Wu, Zeren Jiang, Martin R. Oswald, and Jie Song. From rays to projections: Better inputs for feed-forward view synthesis. *arXiv preprint arXiv:2601.05116*, 2026b. URL https://arxiv.org/abs/2601.05116.

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. *arXiv preprint arXiv:2604.18747*, 2026. URL https://arxiv.org/abs/2604.18747.

Minghao Yin, Jiahao Lu, Wenbo Hu, Wang Zhao, Shan Ying, and Kai Han. SCoPE: Sightline-coordinate positional encoding for video diffusion transformers. *arXiv preprint arXiv:2606.27345*, 2026. URL https://arxiv.org/abs/2606.27345.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a. URL https://arxiv.org/abs/2512.07237.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. GS-LRM: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2404.19702.

Yifan Zhang, Zixiang Chen, Yifeng Liu, Zhen Qin, Huizhuo Yuan, Kangping Xu, Yang Yuan, Quanquan Gu, and Andrew Chi-Chih Yao. Group representational position encoding. In *International Conference on Learning Representations*, 2026b. URL https://arxiv.org/abs/2512.07805.

A Algebra and Proofs

This section separates RDA's exact properties from the assumptions and finite-spectral approximation used by the spectral ablations. The proofs follow the forward pass: first A_i, then the score and content paths, and finally the complete attention computation.

A.1 World-frame changes and guarantee levels

Token transform. Let $L_i \in \mathbb{R}^{3\times 3}$ be nonsingular and define

$$A_i(\alpha) = \begin{bmatrix} L_i & \alpha t_i \\ 0 & 1 \end{bmatrix}, \qquad A_i = A_i(1). \tag{10}$$

For RDA, $L_i = R_i^{\mathrm{cam}} E_i$, where $E_i = [d_{i,x}, d_{i,y}, b_i]$. Nonsingularity of L_i is the geometric assumption required by the affine inverse and content proofs.

Common world gauge. A rigid relabeling of world coordinates is

$$H = \begin{bmatrix} Q & h \\ 0 & 1 \end{bmatrix}, \qquad Q \in \mathrm{SO}(3), \tag{11}$$

and acts on all tokens by $A_i' = HA_i$. For the score path, the scaled matrix obeys

$$A_i'(\alpha) = H_\alpha A_i(\alpha), \qquad H_\alpha = \begin{bmatrix} Q & \alpha h \\ 0 & 1 \end{bmatrix}. \tag{12}$$

Image coordinates p_i remain unchanged because this operation only relabels the world frame.

Guarantee levels. RDA uses two exact algebraic properties, up to floating-point error: common left factors cancel from relative Q/K scores, and the V/O transform has an exact inverse. The spectral ablations add a conditional column isometry that requires a full-rank Gram matrix with an inactive numerical clamp. Their continuous rotation–translation interpretation is otherwise a finite quadrature and truncation approximation.

A.2 Score path

Lemma 1 (dual/primal pairing). For any invertible homogeneous affine matrix A, a dual query and primal key satisfy

$$(A^{-T} q)^T (Ak) = q^T k. \tag{13}$$

This follows directly from $(A^{-T})^T A = A^{-1} A = I$.

Proposition 1 (relative score form). The RDA score between tokens i and j depends on the cameras only through the relative matrix $A_i(\alpha_\ell)^{-1} A_j(\alpha_\ell)$:

$$\widehat{q}_i^T \widehat{k}_j = q_i^T \left(I_{36} \otimes A_i(\alpha_\ell)^{-1} A_j(\alpha_\ell) \right) k_j. \tag{14}$$

Proof. Substitute Eq. (5) and apply the transpose and mixed-product rules for Kronecker products.

Proposition 2 (common-gauge score invariance). Every RDA logit is unchanged by a common rigid world gauge, including the learned translation scale. *Proof.* Equation (12) gives

$$A_i'(\alpha_\ell)^{-1} A_j'(\alpha_\ell) = A_i(\alpha_\ell)^{-1} H_{\alpha_\ell}^{-1} H_{\alpha_\ell} A_j(\alpha_\ell) \tag{15}$$

$$= A_i(\alpha_\ell)^{-1} A_j(\alpha_\ell). \tag{16}$$

Applying this identity to every four-dimensional block proves invariance of the full 144D score and hence of the softmax weights.

A.3 Content path

Lemma 2 (inverse-transpose block). Block inversion gives

$$A_i^{-1} = \begin{bmatrix} L_i^{-1} & -L_i^{-1} t_i \\ 0 & 1 \end{bmatrix}, \qquad A_i^{-T} = \begin{bmatrix} L_i^{-T} & 0 \\ -t_i^T L_i^{-T} & 1 \end{bmatrix}. \tag{17}$$

Direct multiplication yields $A_i^T A_i^{-T} = I_4$, so A_i^T is the exact inverse of the value transform.

Table 6: Score-construction ablation on RealEstate10K at 20,000 updates. Entries are mean ± sample standard deviation over three seeds; the full differential and content path are fixed, and the best metric in each column is shaded.

Configuration	Score dim.	PSNR ↑	SSIM ↑	LPIPS ↓
RDA affine	144	21.882 ± 0.129	0.6847 ± 0.0039	0.2162 ± 0.0010
Fixed spectral	168	21.847 ± 0.158	0.6844 ± 0.0018	0.2187 ± 0.0020
Calibrated spectral	168	21.848 ± 0.065	0.6842 ± 0.0011	0.2194 ± 0.0022

Proposition 3 (local content round trip). The 144D content map in Eq. (7) has the inverse in Eq. (8). *Proof.* Every split-RoPE block is a planar rotation and therefore $\mathcal{R}(p)^{-1} = \mathcal{R}(p)^T = \mathcal{R}(-p)$. The direct-sum product then separates into

$$C_i^{-1} C_i = (I_{18} \otimes A_i^T A_i^{-T}) \oplus I_{36} \oplus I_{36} = I_{144}. \tag{18}$$

Proposition 4 (common-gauge content invariance). The content path is unchanged by a common invertible affine world gauge. Define

$$G_H = (I_{18} \otimes H^{-T}) \oplus I_{72}. \tag{19}$$

Since $(HA_i)^{-T} = H^{-T} A_i^{-T}$ and p_i is unchanged, $C_i' = G_H C_i$ and $C_i'^{-1} = C_i^{-1} G_H^{-1}$. Therefore

$$C_i'^{-1} C_j' = C_i^{-1} G_H^{-1} G_H C_j = C_i^{-1} C_j. \tag{20}$$

For any fixed scalar weights w_{ij}, linearity also gives

$$C_i'^{-1} \sum_j w_{ij} C_j' v_j = C_i^{-1} \sum_j w_{ij} C_j v_j. \tag{21}$$

A.4 Complete attention

Theorem 1 (gauge compatibility of RDA). Under a common rigid relabeling of all cameras, the complete RDA output in Eq. (9) is unchanged before floating-point error. *Proof.* Proposition 2 leaves every attention logit and softmax weight unchanged. Proposition 4 leaves the corresponding weighted relative content computation unchanged. Substituting both results into Eq. (9) proves the claim.

The theorem establishes gauge compatibility for the camera-dependent part of attention; image coordinates, camera intrinsics, and nonlinear host components remain fixed.

A.5 Spectral score ablations and finite-spectral approximation

The two spectral ablations factor the ray differential into an oriented frame and a local calibration matrix. Define

$$F_i = [\widehat{d}_{i,x},\, b_i \times \widehat{d}_{i,x},\, b_i] \in \mathrm{SO}(3), \quad O_i = R_i^{\mathrm{cam}} F_i, \quad T_i = F_i^T E_i, \tag{22}$$

so that $L_i = O_i T_i$. Let $N = \{n_s\}_{s=1}^6$ be antipodal octahedral nodes and $S \in \mathbb{R}^{6\times 4}$ synthesize a real $\ell \le 1$ basis. Both ablations allocate 96 learned dimensions to the exact affine score and 48 learned dimensions to six complex four-dimensional coefficient blocks. Realification lifts the spectral part to 72 score channels, producing a 168D Q/K comparison space.

Table 6 keeps the full differential and RDA content path fixed while replacing the 144D affine score with two 168D spectral alternatives. Neither wider construction improves the three-seed mean. The calibrated spectral variant is 0.034 dB lower in PSNR, 0.0005 lower in SSIM, and 0.0032 higher in LPIPS than the affine score. Additional score capacity and spectral structure are therefore unnecessary for the observed RDA gain.

Fixed spectral transform (internal M2). The fixed ablation uses the same low-band synthesis S for every token, rotates it by the world-oriented patch frame, and applies translation phase at the chosen radial frequencies. It tests whether a fixed finite spectral branch helps beyond RDA's affine score.

Table 7: Guarantee ledger for RDA and its score ablations. "Exact" denotes an algebraic identity before floating-point error; whitening isometry additionally assumes a full-rank, unclamped Gram matrix.

Variant	Q/K construction	Exact or conditional result	Spectral approximation
RDA	144D affine transform	Exact score invariance and V/O inverse; exact content invariance	None beyond numerical inversion
Fixed spectral	96D affine plus 72D fixed spectral	RDA affine/content results; exact retained-band rotation	Combined spectral rotation–translation transform
Calibrated spectral	96D affine plus 72D Galerkin spectral	Fixed-row results; conditional whitening isometry	Combined spectral transform and clamp-active isometry

Calibrated spectral transform (internal M3). The calibrated ablation first forms

$$[\Gamma_i]_{sr} = \exp\left(-\frac{\|T_i^T n_s - n_r\|^2}{2\sigma^2}\right), \qquad B_i = \Gamma_i S, \tag{23}$$

and then right-whitens the active columns:

$$G_i = B_i^T B_i, \qquad U_i = B_i (G_i)_\epsilon^{-1/2}, \qquad \Phi_i = D(O_i) U_i. \tag{24}$$

For radial frequency ρ_b, its spectral lift is

$$\mathcal{S}_{i,b}(c) = \mathrm{diag}\big(e^{-\mathrm{i}\alpha_\ell \rho_b N t_i}\big)\, \Phi_i c. \tag{25}$$

Lemma 3 (whitening preserves the active subspace). Let $B \in \mathbb{R}^{6\times 4}$ have full column rank, $G = B^T B$, and $U = BG^{-1/2}$. If the eigenvalue clamp is inactive, then

$$U^T U = G^{-1/2} G G^{-1/2} = I_4. \tag{26}$$

Because $G^{-1/2}$ is invertible, $\mathrm{col}(U) = \mathrm{col}(B)$. Whitening changes coordinates inside the active columns but cannot create their token-dependent ambient subspace; that dependence enters through $B_i = \Gamma_i S$. If the clamp is active, the right factor still preserves the column space but need not remain isometric.

Finite-spectral approximation. The sampled low-band rotation is exact on its retained subspace, and each diagonal translation phase is unitary. Their combination approximates the full continuous $\mathrm{SE}(3)$ transform because translation generates angular frequencies outside the retained band and six samples are not closed under arbitrary rotations. In general,

$$D(Q)\mathcal{T}_\rho(t)D(Q)^{-1} \neq \mathcal{T}_\rho(Qt) \tag{27}$$

for the truncated matrices. Both ablations retain RDA's exact affine score branch and content path; spectral equivariance follows the measured quadrature/truncation approximation.

B Implementation Specifications

B.1 RDA implementation

The ray differential E_i is constructed analytically from K_i^{-1} at pixel-center patch coordinates. The $W/2$ and $H/2$ factors in Eq. (2) convert pixel derivatives to normalized image-coordinate derivatives. Geometry and cached linear inverses are computed in FP32 before features are cast back to the model dtype.

The 144D score head contains 36 affine blocks. The score translation scale is learned independently in each of the six attention layers from 0.01. The content transform contains 18 inverse-transpose blocks (72D), 36D image-x RoPE, and 36D image-y RoPE. Image RoPE uses patch position $p_i = \mathrm{patch}_i/8 - 0.5$, base 100, and unit frequency scale; content translation is fixed to one.

Table 8: Per-seed RealEstate10K results on the fixed primary panel. Each cell is the scene mean for one completed 20k checkpoint. The best result within each seed and metric is shaded light green.

	Seed 0			Seed 1			Seed 2		
Method	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
CaPE	20.822	0.6415	0.2613	20.526	0.6262	0.2754	20.580	0.6321	0.2652
PRoPE	21.312	0.6710	0.2399	21.299	0.6665	0.2405	21.418	0.6705	0.2220
GTA	21.137	0.6543	0.2517	21.247	0.6572	0.2490	21.171	0.6577	0.2456
URoPE	20.238	0.6302	0.2712	20.274	0.6272	0.2790	20.165	0.6243	0.2616
UCPE	21.230	0.6736	0.2225	21.596	0.6754	0.2249	21.483	0.6709	0.2259
RDA	21.734	0.6802	0.2159	21.938	0.6867	0.2173	21.974	0.6872	0.2153
Fixed spectral	21.665	0.6824	0.2169	21.953	0.6852	0.2208	21.924	0.6857	0.2186
Calibrated spectral	21.781	0.6829	0.2176	21.852	0.6845	0.2218	21.911	0.6851	0.2188

Table 9: Per-seed DL3DV results on the fixed 140-scene test panel. Each cell is the scene mean for one completed 100,000-update checkpoint. The best result within each seed and metric is shaded light green.

	Seed 0			Seed 1			Seed 2		
Method	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
CaPE	14.3757	0.3449	0.8111	14.5603	0.3530	0.7680	14.5388	0.3521	0.7466
PRoPE	14.6316	0.3575	0.7299	14.6325	0.3581	0.7303	14.6442	0.3584	0.7374
GTA	14.6593	0.3577	0.7229	14.6159	0.3559	0.7331	14.6363	0.3571	0.7314
URoPE	14.4941	0.3541	0.7777	14.4344	0.3522	0.7864	14.4925	0.3547	0.7676
UCPE	14.7053	0.3584	0.7350	14.6915	0.3584	0.7383	14.7082	0.3586	0.7392
RDA	14.7085	0.3608	0.7175	14.7133	0.3609	0.7085	14.7178	0.3606	0.7102

B.2 Spectral-ablation implementation

The two ablations use six octahedral directions, real $\ell \leq 1$ synthesis, and radial frequencies $(0.5, 1, 0.5, 1, 0.5, 1)$. The calibrated row uses Gaussian bandwidth 0.35 and clamps Gram eigenvalues at 10^{-5}. Both keep the same per-layer score translation scalar and the same RDA content transform.

The internal experiment identifiers are retained only for correspondence with existing configurations and checkpoints:

Internal ID	Paper role	Score construction
M1	RDA, proposed method	144D affine transform
M2	Fixed spectral ablation	96D affine + 72D fixed spectral
M3	Calibrated spectral ablation	96D affine + 72D Galerkin spectral

C Per-Seed Primary Results

Table 8 retains the individual test means used to construct Table 1 and Table 6. Aggregation is performed across the three seed-level means, not across pooled per-scene predictions, so training-run variation remains distinct from within-test-set variation.

D Per-Seed DL3DV Results

Table 9 retains the individual scene means used to construct the aggregated DL3DV result in Table 2. Each entry evaluates the completed 100,000-update checkpoint on the fixed 140-scene test manifest. The best result within each seed and metric is shaded light green; aggregation is performed across these seed-level means.

E Efficiency Analysis and Profiling

RDA preserves the host's 144D score width, so every method in Table 10 has the same dominant dense-operation count of 1.095 TFLOPs. Their realized costs differ because the camera transforms and memory access patterns are not included in this leading-order count. RDA takes 59.60 ms, closely matching GTA at 60.24 ms and remaining faster than UCPE at 91.57 ms. Its 310 MiB incremental memory is below URoPE's 349 MiB but above the other rows. RDA therefore obtains its reconstruction gains without increasing score width or dominant FLOPs, although it is neither the lowest-latency nor the lowest-memory method.

Table 10: Same-device cost of a complete six-layer LVSM forward with batch size two, two context views, one target view, 3,072 tokens, and FP16 on an RTX 4090. Latency is the median of 20 runs after five warmups; peak memory is incremental allocation above the resident model and inputs. CaPE† modifies Q/K only, while all other rows also transform V/O.

Method	FLOPs (TFLOPs) ↓	Latency (ms) ↓	Peak memory (MiB) ↓
CaPE†	1.095	28.56 (1.00×)	227
PRoPE	1.095	46.49 (1.63×)	255
GTA	1.095	60.24 (2.11×)	255
URoPE	1.095	29.32 (1.03×)	349
UCPE	1.095	91.57 (3.21×)	255
RDA	1.095	59.60 (2.09×)	310

For batch size $B = 2$, $H = 8$ heads, $N = 3072$ tokens, value width $D_v = 144$, and score width D_{score}, the two dominant attention matrix products require

$$\text{FLOPs}_{\text{attn}} = 2BHN^2\big(D_{\text{score}} + D_v\big), \tag{28}$$

where one multiplication and one addition count as two FLOPs. This is 86.97 GFLOPs per layer at score width 144. Adding tokenization, Q/K/V/O projections, six feed-forward blocks, and the output projection gives the 1.095 TFLOP full-host count in Table 10. This count omits lower-order softmax, normalization, trigonometric, and small-matrix operations; measured latency captures their realized cost.

Table 11: Detailed same-device efficiency profile. Single-layer F+B is the median of eight repetitions after three warmups of the steady-state geometry-aware SDPA bridge; packet-specific caches are warm, and common packet construction and host projections are excluded. Full throughput includes tokenization, geometry-packet construction, all six Transformer layers, and RGB output projection.

Method	Q/K width	Attn. GFLOPs/layer	Attn. F+B (ms) ↓	Full scenes/s ↑
CaPE†	144	86.97	3.65	70.0
PRoPE	144	86.97	11.67	43.0
GTA	144	86.97	13.07	33.2
URoPE	144	86.97	7.07	68.2
UCPE	144	86.97	14.35	21.8
RDA	144	86.97	7.27	33.6

The measurements use PyTorch 2.5.1, FP16, Flash SDPA, and an otherwise idle RTX 4090. Full throughput is $2{,}000$ divided by the median batch-two latency from 20 timed forwards after five warmups. CaPE provides the Q/K-only latency reference. URoPE has the largest memory footprint because its source-exact path expands keys and values across query views and precomputes projected depth-anchor coordinates.

F Experimental Protocol Details

The following settings reproduce the reported study.

- *Training.* Every row uses the same six-layer, eight-head LVSM host, 144D learned head width, and FP16 training. The RealEstate10K primary comparison uses the predeclared 20,000-update endpoint, whereas the DL3DV rows use the predeclared 100,000-update endpoint. Both studies use seeds 0, 1, and 2; methods within a seed share the host initialization and scene order.
- *Checkpoint selection.* Evaluation uses the protocol-defined step-20,000 or step-100,000 checkpoint rather than validation performance. Missing, unexpected, or shape-mismatched parameters invalidate a row.
- *Evaluation population.* The RealEstate10K primary panel uses two context views, three targets, and 256×256 images. The DL3DV test panel contains 140 test scenes, two context views, ten targets, a 100-frame span, 256×256 images, and equal scene weighting.
- *Camera intervention.* The crop audit fixes one ordered panel and applies six paired RGB-plus-intrinsics transformations: centered crops at 0.875 and 0.75 and four one-sided crops at 0.875. Each crop is resized to 256×256 with the matching focal-length and principal-point update.
- *Artifacts.* We retain training configurations, evaluation manifests, strict-load diagnostics, and per-scene PSNR/SSIM/LPIPS records for every reported run.