ATLASFLOW: SEAM-AWARE SAMPLING AND PERSPECTIVE DISTILLATION FOR PANORAMA GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

An equirectangular panorama is stored with an arbitrary longitude cut, and planar diffusion priors attach meaning to it: released ERP generators break there, and their samples change when the cut moves. We remove this dependence without updating the panorama backbone. AtlasFlow evaluates the frozen denoiser in two opposite longitude charts, aligns the predicted velocities, and fuses them with a symmetric partition of unity that downweights each chart near its cut. The sampler is exactly half-turn equivariant in latent space; co-rotating the chart layout extends this to every integer roll, randomized per sample at no extra cost. On Matterport3D the stored-cut error of matched DiT360 falls by 31% at no cost in aggregate quality: the complete method leads eight of ten reference and quality metrics among twelve generators and improves eight over DiT360; on ODI-SR1200 it reduces FID from 59.64 to 55.02 and FAED from 16.82 to 14.64. An 18.1K-parameter Perspective Velocity Adapter (PVA), a periodic residual distilled from the denoiser's own square-view predictions with no teacher at inference, composes with the sampler and supplies its reference-quality and viewport gains. We further show that frame-boundary seam metrics miss a relocated seam and depend strongly on decoding: a lossless half-turn roll passes them, whereas a yaw-invariant prominence read-out detects injected off-boundary seams with AUC 0.92 where the boundary metric stays at chance.

1 INTRODUCTION

An equirectangular projection (ERP) stores a full viewing sphere in one rectangular image. Its left and right borders are adjacent rays in the scene but separated boundaries in the array; if they disagree, wrapping the image around a viewer exposes a vertical crack. The location of this boundary is arbitrary: changing the zero of longitude changes only the storage chart, not the physical panorama, so a generator should attach no significance to a particular cut.

Recent methods adapt perspective-image diffusion priors to panoramas through multi-view correspondence, projection-aware branches, spherical operators, or hybrid perspective–panorama training (Tang et al., 2023; Zhang et al., 2024; Sun et al., 2025; Feng et al., 2026). Direct ERP generation keeps one global latent and scales to high resolution, but it inherits two mismatches. First, the denoising trajectory is *cut dependent*: changing the longitude origin changes the sample even after its output is unrolled to the original frame. DiT360's duplicate token padding exposes cross-boundary context but is not a transformation constraint, and its released sampler retains the cut dependence measured below. Second, FLUX is pretrained on perspective images, so the velocity predicted for a local ERP region can disagree with the same frozen model's prediction for that region in its familiar square coordinates (Figure 1).

We address both failures in the velocity field rather than in decoded pixels. AtlasFlow, our training-free sampler, expresses the current ERP latent at each step in two opposite charts, evaluates the same frozen model in each, undoes both transforms, and combines the aligned velocities with a symmetric partition of unity before one shared scheduler update. Sinusoidal weights give exact C_2 equivariance at the packed latent for any base denoiser; co-transporting the chart layout extends this to every integer longitude roll, and one random layout per sample gives the distributional form. The

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



Figure 1: **The two failures, seen through a $90°$ camera.** Left: turned to the stored ERP boundary (marked), a bed's headboard ends where a door begins; the two halves of the room do not meet across the cut. Right: pitched $45°$ down, a straight kitchen counter bends into an arc. Outputs of released ERP generators for Matterport3D prompts.

Perspective Velocity Adapter (PVA), an 18,128-parameter periodic residual, then corrects the fused prediction using the frozen model's own square-view velocity as a training-only teacher.

This paper makes three contributions.

- **AtlasFlow**, a training-free sampler that makes a frozen ERP diffusion model exactly equivariant to the storage cut: two opposite charts, one denoiser call each, fused by a symmetric partition of unity before one shared update. It is the closed-form solution of a weighted consensus problem of which uniform window averaging is a special case; it removes DiT360's cut break (SeamL1 -31%, SeamGrad -38%) at parity or better on the aggregate metrics, closes frozen FLUX without any panorama training (DS $71 \to 3$), and extends by layout co-transport to every integer roll, with a randomized form that is equivariant in distribution.
- A seam-measurement finding and a fix: frame-boundary metrics are passed by relocating the seam and swing with the decoder, whereas a yaw-invariant prominence read-out (RingCut-P) detects injected off-boundary seams with AUC 0.92 where the boundary metric is at chance.
- **PVA**, an 18K-parameter periodic residual distilled from the same frozen model's square-view predictions, with no teacher at inference; composed with AtlasFlow it improves FID $37.49 \to 36.30$, FAED $3.47 \to 3.11$, and BRISQUE $22.96 \to 20.99$ over the sampler alone and the depth consistency of rendered views.

2 Related Work

Text-to-panorama generation. MVDiffusion, PanFusion, and SMGD generate overlapping perspective views, couple perspective and panorama branches, or introduce spherical-manifold operators (Tang et al., 2023; Zhang et al., 2024; Sun et al., 2025); StitchDiffusion and UniPano revisit a stitch region or adapt LoRA weights (Wang et al., 2024a; Ni et al., 2025). DiT360 hybrid-trains a FLUX-style transformer on perspective and polished panoramic data; its released model is our frozen base (Feng et al., 2026). TanDiT generates on a tangent-plane grid (Çapuk et al., 2025), whereas our state remains one ERP latent and tangent views enter only the PVA training objective. SpheRoPE and SHERPA modify positional encoding or internal token representations (Hirschorn et al., 2026; Kang et al., 2026); AtlasFlow leaves the released positional encoding and text condition unchanged and acts on aligned full-ERP velocity predictions, and PVA learns a residual on the denoiser's output rather than on hidden features.

Sampling-time consistency and equivariance. MultiDiffusion and SyncDiffusion fuse overlapping denoising windows, and SpotDiffusion shifts non-overlapping windows over time (Bar-Tal et al., 2023; Lee et al., 2023; Frolov et al., 2025); PanoDiffusion and 360DVD rotate latents over time, and Diffusion360 blends horizontal boundaries after each update (Wu et al., 2024; Wang et al., 2024b; Feng et al., 2023). AtlasFlow instead evaluates alternate full-ERP charts *at the same timestep*, aligns their complete predictions, and fuses them *before* one shared update; it neither maintains separate trajectories nor overwrites a boundary afterwards. Generic symmetrization imposes equivariance by group averaging or transformed inverse maps (Zhang et al., 2025; Wu et al., 2026a); AtlasFlow specializes it to ERP cuts by assigning responsibility by distance from each artificial boundary, with uniform averaging as the control that separates algebraic symmetry from cut reliability.

Perspective priors and panorama assessment. Latent diffusion models and diffusion transformers inherit strong local image priors from perspective-image training (Rombach et al., 2022; Peebles & Xie, 2023; Black Forest Labs, 2024); PVA uses the same frozen panorama-capable denoiser as

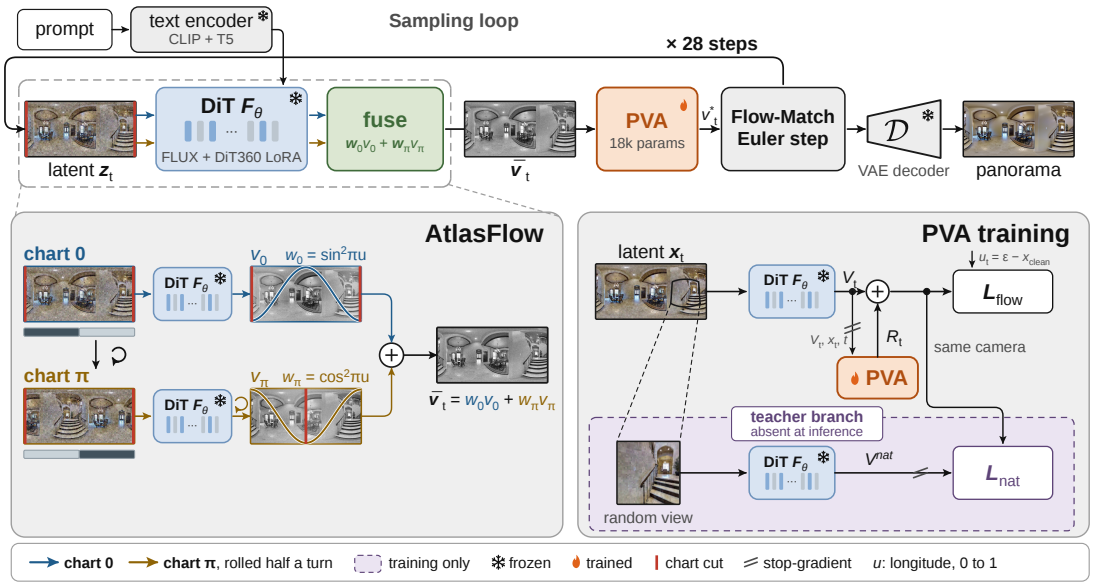

Figure 2: **The fixed-layout pipeline: AtlasFlow sampling with PVA.** Top row: one of the 28 denoising steps. Inside the dashed frame the same noisy ERP latent is read in two longitude charts, the one frozen DiT360 denoiser F_θ is called once per chart, and *fuse* combines the aligned velocity fields with the symmetric partition of unity of Eq. (6); PVA adds one periodic residual and a single FlowMatch Euler update advances the shared latent. Lower left, **AtlasFlow**: chart π is the same latent rolled by half a turn (the gold branch carries that half-turn in and back out), so each chart's cut lies in the other's interior and the weight drawn over each velocity field falls to zero at its own cut. Lower right, **PVA training**: one random 90° tangent camera per minibatch gives a stopped teacher prediction V^{nat} from the same frozen denoiser run on the crop as a plain square image (F_θ^{sq}); the corrected ERP velocity is projected with the same camera (Π_i^v) for the student loss. Only PVA is trained; the teacher branch is absent at inference.

student backbone and training-only teacher, distilling their local velocity difference without a separate teacher model or RGB input at inference. RL-ScanIQA and PanoDecouple's Distort-FID assess panorama quality and distortion-aware distribution distance (Wang et al., 2026; Zheng et al., 2025); our geometry probes instead use Perspective Fields and Depth Anything V2 (Jin et al., 2023; Yang et al., 2024a) as frozen probes of camera plausibility and overlapping-view structure, neither of which supplies the PVA training target. Extended related work is given in Appendix I.

3 Problem Formulation

Let $x_t \in \mathbb{R}^{C\times H\times W}$ be a noisy VAE latent, $z_t = \mathrm{Pack}(x_t) \in \mathbb{R}^{H_p\times W_p\times D}$ its packed ERP tokens, and c a text condition. A released FLUX-style panorama model predicts a packed rectified-flow velocity $F_\theta(z_t, t, c, q)$, where q denotes image-token coordinates. We keep θ fixed and ask for two properties: the generated sphere should not depend on where longitude is cut for storage, and each rendered tangent view should retain plausible local geometry.

Cut sensitivity under longitude reparameterization. Let ρ_g cyclically roll longitude by chart offset g, P add periodic ghost columns, and C remove their predictions. The physical-frame velocity predicted through chart g is

$$v_g(z_t, t, c) = \rho_g^{-1} C F_\theta(P(\rho_g z_t), t, c, q). \tag{1}$$

The coordinate convention q is held fixed: charting permutes the latent content presented to the model and unrolls its prediction, while token IDs stay attached to chart-grid positions. If S_θ denotes the matched sampler from initial noise to final packed latent, its cut sensitivity is

$$\mathcal{D}_{\mathrm{cut}}(g) = \frac{\|\rho_g^{-1} S_\theta(\rho_g z_T) - S_\theta(z_T)\|_2}{\|S_\theta(z_T)\|_2}, \tag{2}$$

which should vanish for the selected longitude transformations. The released sampler, with its duplicate-ID token padding, has mean $\mathcal{D}_{\mathrm{cut}}(\pi) = 0.3447$ over eight fixed prompts, a deterministic property of the sampler, not a statistical estimate (Table 7).

Perspective-compatible local velocity. Let Π_i^z and Π_i^v bilinearly sample the same tangent camera i from an ERP latent and velocity, and let F_θ^{sq} denote the frozen denoiser evaluated with square-grid coordinates, without PVA or ERP token padding but retaining the DiT360 LoRA. For a candidate ERP velocity V, the native-view discrepancy is

$$\mathcal{D}_{\mathrm{view}}(V) = \mathbb{E}_i \Big[w(t) \, \| \Pi_i^v V - \mathrm{sg}(F_\theta^{\mathrm{sq}}(\Pi_i^z x_t, t, c)) \|_2^2 \Big], \tag{3}$$

where sg stops gradients. Both sides share prompt, timestep, noisy tangent latent, and weights; they differ in the coordinate regime. The goal is a physical velocity that controls $\mathcal{D}_{\mathrm{cut}}$ while reducing $\mathcal{D}_{\mathrm{view}}$ without replacing the backbone.

4 METHOD

Figure 2 follows one denoising step. AtlasFlow (Sec. 4.1) evaluates the frozen denoiser under two longitude charts, maps every prediction back to one physical frame, and fuses the aligned velocities; PVA (Sec. 4.2) then applies one periodic correction before the shared scheduler update. Unless stated otherwise, *AtlasFlow* denotes the two-chart symmetric partition-of-unity (PoU) sampler of Eq. (6) with a fixed layout, and *Ours* denotes the complete configuration: AtlasFlow followed by the frozen PVA trained with the native square-view teacher.

4.1 SEAMLESS LONGITUDE CLOSURE VIA CUT-RELIABILITY FUSION

A single ERP chart forces a closed sphere to have one artificial boundary. Rolling the chart merely relocates which region is exposed to it, and merging decoded images acts after the trajectories have diverged. AtlasFlow instead queries alternative charts at the same time step and combines their complete velocity fields before any update, trusting each chart least near its own cut. For a finite chart set G with nonnegative reliabilities $w_g(x)$ summing to one, the fused physical velocity is

$$\bar{v}(z_t, t, c)(x) = \sum_{g \in G} w_g(x) v_g(z_t, t, c)(x), \tag{4}$$

which is also the closed-form solution of a local weighted consensus problem,

$$\bar{v}(x) = \arg\min_u \sum_{g \in G} r_g(x) \|u - v_g(x)\|_2^2, \qquad w_g(x) = \frac{r_g(x)}{\sum_{h \in G} r_h(x)}. \tag{5}$$

The weights therefore encode an explicit reliability assumption rather than an unstructured ensemble. Only one fused field leaves the multi-chart branch, so all charts share one latent trajectory and one scheduler state. For two opposite charts on packed-token centers $x + 1/2$ we use

$$w_0(x) = \sin^2\left(\frac{\pi(x + 1/2)}{W_p}\right), \qquad w_\pi(x) = \cos^2\left(\frac{\pi(x + 1/2)}{W_p}\right), \tag{6}$$

which sum to one, exchange under a half-turn, and downweight each chart near its own cut in favor of the chart whose cut is half a turn away.

Proposition 1 (Finite-chart equivariance)**.** *If $G = C_K$ is complete and $w_g(\rho_h x) = w_{gh}(x)$, then $\bar{v}(\rho_h z, t, c) = \rho_h \bar{v}(z, t, c)$ for every $h \in C_K$, and an elementwise scheduler preserves the relation along the whole trajectory.*

The proof (Appendix J) re-indexes the finite sum after Eq. (1) rolls every prediction and permutes its chart index; it makes no equivariance assumption on F_θ. Uniform $1/K$ averaging also satisfies the proposition, which separates algebraic symmetry from cut reliability; the latter is tested directly in Sec. 5.2. The two-chart implementation has zero packed-latent half-turn error on all test cases but guarantees neither quarter-turn nor continuous $SO(2)$ equivariance. The PVA below contains only horizontally circular convolutions, pointwise operations, and roll-invariant normalization, so it commutes with integer rolls and the complete method inherits the guarantee.

The layout as a sampler coordinate. The guarantee above concerns a subgroup of shifts. Let R_a roll by a packed columns, let $L = (a_1, \ldots, a_K)$ denote the cut positions, and let $V_L(z, t, c) = \sum_i w_i^L \odot R_{a_i} \widetilde{F}(R_{-a_i} z, t, c)$ be the fused field with $\widetilde{F}$ the frozen denoiser including padding and cropping. Moving each weight map with its cut gives $w_i^{L+b} = R_b w_i^L$, and direct substitution yields

$$V_{L+b}(R_b z, t, c) = R_b V_L(z, t, c), \qquad b \in \mathbb{Z}_{W_p}, \tag{7}$$

for either weight profile and any backbone: joint input–layout equivariance for all integer rolls (Appendix K). Drawing δ uniformly once per sample and keeping $L + \delta$ fixed along the trajectory gives equivariance in distribution at no extra denoiser calls. This is a random-rotation wrapper, not a per-sample guarantee under a fixed drawn layout, and by itself it removes no defect: with one chart it merely relocates the seam, the loophole Sec. 5.6 quantifies.

4.2 Perspective Velocity Adapter

PVA addresses Eq. (3) by comparing the student's ERP velocity on a random tangent crop with the velocity the same frozen denoiser predicts when that noisy crop is presented as an ordinary square image. It operates on the unpacked 16-channel VAE velocity map: a 1×1 projection maps the stopped base velocity, noisy latent, and a constant timestep plane (33 channels) to 64 hidden channels, followed by three depthwise-separable 3×3 residual blocks with dilations $1, 2, 4$ and a zero-initialized 1×1 output projection, for 18,128 parameters. Horizontal padding is circular and vertical padding replicates the boundary row; the adapter has no absolute-longitude input, so A_ϕ commutes with cyclic rolls.

Let V_t be the unpacked frozen one-chart velocity and $R_t = A_\phi(\mathrm{sg}(V_t), x_t, t)$. For a sampled tangent camera i we extract the noisy square latent $\Pi_i^z x_t$, query the frozen denoiser with square-grid coordinates, $V_{t,i}^{\mathrm{nat}} = F_\theta^{\mathrm{sq}}(\Pi_i^z x_t, t, c)$, and compare it with the student view $V_{t,i}^{\mathrm{stu}} = \Pi_i^v(V_t + R_t)$:

$$\mathcal{L}_{\mathrm{nat}} = \mathbb{E}_i\Big[w(t)\,\big\|V_{t,i}^{\mathrm{stu}} - \mathrm{sg}(V_{t,i}^{\mathrm{nat}})\big\|_2^2\Big]. \tag{8}$$

Teacher and student share content, noise state, prompt, timestep, and weights; the coordinate regime changes, and with it the context and noise statistics the network sees. With rectified-flow target $u_t = \epsilon - x_{\mathrm{clean}}$ (Lipman et al., 2023; Black Forest Labs, 2024), the objective is

$$\mathcal{L}_{\mathrm{flow}} = \mathbb{E}\big[w(t)\|V_t + R_t - u_t\|_2^2\big], \qquad \mathcal{L} = \mathcal{L}_{\mathrm{flow}} + \lambda_{\mathrm{nat}} a_k \mathcal{L}_{\mathrm{nat}}, \tag{9}$$

with $\lambda_{\mathrm{nat}} = 0.10$, a linear warm-up a_k, and one teacher crop per minibatch; all FLUX and DiT360 weights are frozen (Appendix B, including the checkpoint's reflected training-camera convention). At inference, with $\bar{V}_t = \mathrm{Unpack}(\bar{v}_t)$ from Eq. (4), PVA is applied once after fusion,

$$R_t = A_\phi(\mathrm{sg}(\bar{V}_t), x_t, t), \qquad v_t^\star = \mathrm{Pack}(\bar{V}_t + R_t), \qquad z_{t-1} = \mathrm{Step}(z_t, v_t^\star, t). \tag{10}$$

PVA-only substitutes the released one-chart velocity for $\bar{V}_t$; AtlasFlow-only sets $R_t = 0$.

5 Experiments

Setup. Every controlled comparison against DiT360 fixes prompts, per-image seeds, resolution (2048×1024), 28 steps, guidance, base checkpoint, image IDs, scheduler, and decoder policy in a machine-readable manifest. Following MVDiffusion and PanFusion, the reference-based Matterport3D evaluation uses their public test split, about 1k panoramas with released captions as prompts (Chang et al., 2017; Tang et al., 2023; Zhang et al., 2024). ODI-SR1200, the 1,200 panoramas of ODI-SR (Deng et al., 2021) captioned with the public SpheRoPE procedure, is the cross-dataset evaluation (Appendix A).

Metrics. Table 1 reports ten aggregate scores drawn from the DiT360 evaluation protocol (Feng et al., 2026) plus KID (Binkowski et al., 2018), grouped by what they evaluate. Realism against the reference set: Fréchet distances between Inception features of generated and real images (FID on the whole ERP, $\mathrm{FID}_{\mathrm{crop}}$ without the polar 12.5% bands, the released "FIDclip", and $\mathrm{FID}_{\mathrm{pole}}$ and $\mathrm{FID}_{\mathrm{equ}}$ on the polar and equatorial bands alone), KID, the unbiased kernel form of the same comparison, and FAED (Oh et al., 2022), the same distance in the feature space of a panorama autoencoder, computed with PanFusion's RGB implementation. Diversity: IS with a Places365 scene

Table 1: **Reference and quality metrics on Matterport3D.** Gold, silver, and bronze mark first, second, and third place among the displayed rows (raw ranks, not paired significance). All rows use aligned prompt/reference identities and the same metric code; external systems keep their released generation protocols, while DiT360 and Ours also share prompts, seeds, scheduler, token convention, and decoder. Metrics are defined in Sec. 5; KID is $\times 10^3$ (polynomial kernel, 100 subsets of 1,000).

Method	FID ↓	FID_{crop} ↓	FID_{pole} ↓	FID_{equ} ↓	FAED ↓	KID$\times 10^3$ ↓	IS ↑	CS ↑	BRISQUE ↓	NIQE ↓
MVDiffusion	125.69	110.32	445.56	65.27	22.340	99.54	5.663	**0.3482**	53.00	4.596
PanFusion	262.33	270.66	179.99	209.58	18.675	223.98	**7.216**	0.2880	25.96	4.811
Diffusion360	169.22	170.86	140.28	145.30	22.364	135.08	5.263	0.2993	23.55	4.535
WorldGen	51.35	44.87	243.73	50.80	12.422	25.73	5.218	0.3429	37.28	4.764
Matrix-3D	42.89	36.70	281.73	45.50	6.465	15.86	5.461	0.3411	21.25	4.441
LayerPano3D	48.65	44.45	338.95	52.73	11.401	23.07	5.404	0.3411	36.00	4.619
SMGD	47.03	39.23	319.12	57.66	11.489	16.33	4.451	0.3215	26.09	4.759
PAR	48.44	40.06	309.39	54.93	6.013	18.22	5.194	0.3418	27.19	4.340
HunyuanWorld	73.90	70.48	123.90	56.80	21.104	50.42	5.990	0.3382	32.66	5.146
SpheRoPE	91.80	82.33	205.39	58.26	35.550	57.78	7.007	0.3452	33.54	4.600
DiT360	37.90	36.41	62.30	32.30	3.567	10.49	5.359	0.3436	22.93	4.116
Ours	**36.30**	**34.43**	**62.17**	**31.75**	**3.114**	**8.75**	5.261	0.3433	**20.99**	**4.050**

classifier. Prompt adherence: CLIPScore (CS). Perceptual quality with no reference: BRISQUE and NIQE, which score how far local luminance statistics depart from those of natural images. Lower is better except for IS and CS. Three seeds of the complete method move FID by 0.1 and FAED by 0.05 (Appendix L), an order of magnitude below its gaps to DiT360.

None of these scores isolates the cut, so thirteen paired endpoints and a bootstrap win/tie/loss rule were fixed before the reported runs. They are reported for the matched DiT360–Ours pair: a blurred or low-resolution panorama scores well on seam and viewport diagnostics without being better, so across systems these scores would rank artifacts, and the question here is what changes on top of DiT360's high-resolution output. Cut continuity: SeamL1/L2 and SeamGrad compare the first and last ERP column (the latter normalized by interior variation), DS (Christensen et al., 2024) is the context-normalized Scharr discontinuity at the boundary, StripL1 the mean-color difference of the two 16-column boundary strips, and RingCut the DS operator scanned over every cyclic longitude, a yaw-invariant guardrail against improving a fixed-boundary score by merely moving the artifact. Semantics: CLIPScore and three CLIP panorama-likeness scores. Local geometry, on 16 calibrated 90° views per ERP: VPFD, the error in degrees of a frozen Perspective-Fields camera estimate against the known camera, and VDRI/VDAE, rank and affine depth consistency of Depth Anything V2 across overlapping views; lower is better for the seam and viewport endpoints. A win or loss requires the 95% paired bootstrap interval over panoramas to exclude zero. Definitions and calibration are in Appendices E and H.

5.1 Comparison with prior work

Table 1 reports reproduced outputs of ten released systems, from MVDiffusion, PanFusion and Diffusion360 to HunyuanWorld and the training-free SpheRoPE (Zhang et al., 2024; Tang et al., 2023; Feng et al., 2023; Xie, 2025; Yang et al., 2025; 2024b; Sun et al., 2025; Wang et al., 2025; HunyuanWorld Team, 2025; Hirschorn et al., 2026), each run with its released model and inference configuration; this measures end-to-end output quality, whereas the DiT360–Ours pair is the matched controlled comparison. Parity with DiT360 on these aggregate scores is the expected outcome, since backbone, weights, and decoder are frozen; the finding is that the cut repair costs nothing there and moves eight of the ten columns the right way: Ours improves and leads all four FIDs, FAED, KID, BRISQUE, and NIQE. Each of three seeds improves seven of the nine metrics re-evaluated per seed (Appendix L).

5.2 Continuity at the stored cut

Visual evidence. Figure 3 shows the cut itself. DiT360 and Ours share backbone, prompt, seed, scheduler, and decoder, so the difference at the cut is the sampler: DiT360's windows, frames, and stairs break where its boundary columns meet; Ours is continuous there. PAR and SpheRoPE are context: their blur, warping, and duplication come from their own backbones; Ours keeps the

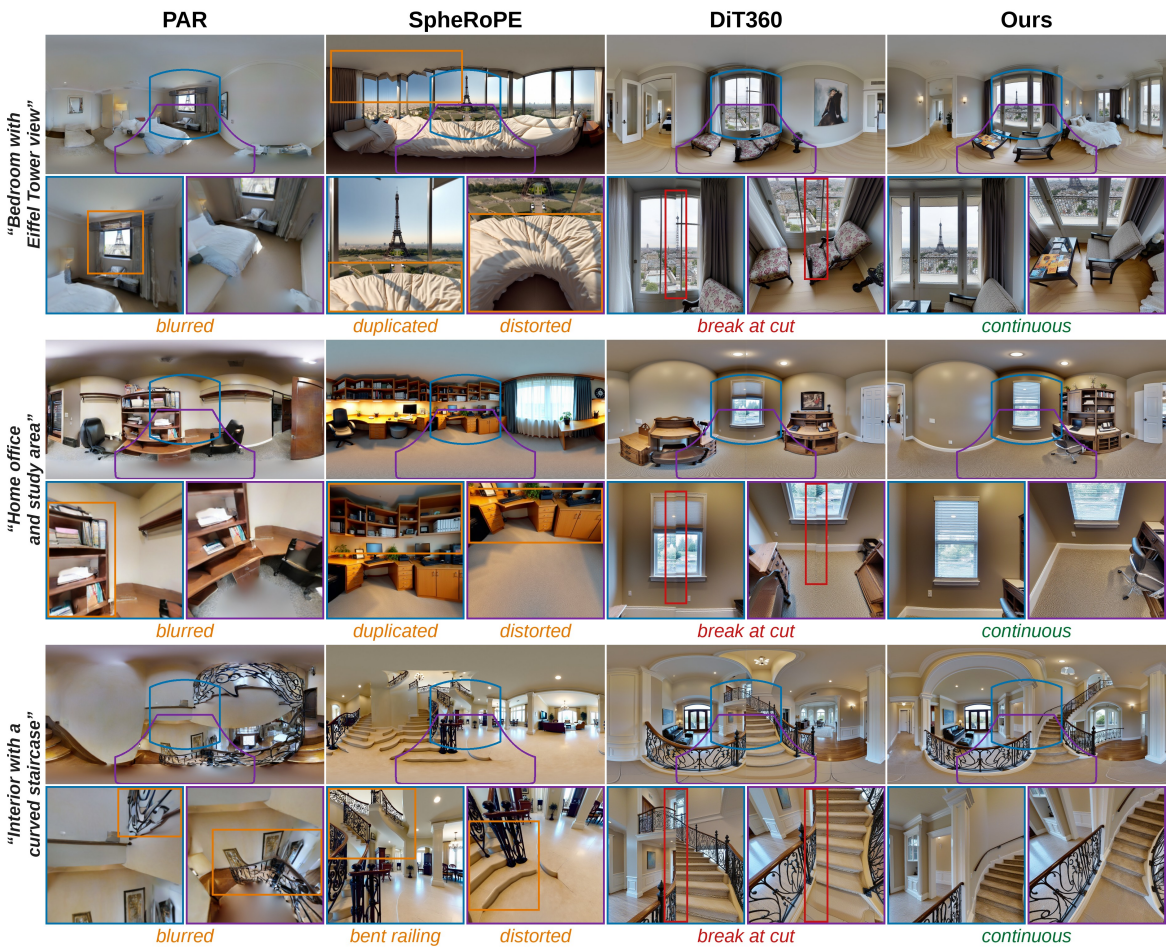

Figure 3: **Qualitative comparison on Matterport3D cohort prompts** (PAR at its native 1024×512). ERPs are rolled so that the stored cut is at the centre; below, two square $90°$ viewport-protocol cameras at the cut longitude, level (**blue**) and $45°$ down (**purple**); PAR, first row, is aimed at the tower window so that all four cells match. **Red**: a break at the cut, which AtlasFlow removes. **Orange**: blur, distorted geometry, and duplicated pattern, which the FLUX/DiT360 prior avoids and Ours inherits. More in Appendix O.

FLUX/DiT360 prior's freedom from them. Figure 4(a) shows the leave-one-out pair: split at the cut with PVA alone, one continuous room with the complete method.

Metric evidence. Table 2, the complete 2×2 leave-component-out design on the cohort (*w/o PVA and AtlasFlow* is released DiT360, *w/o PVA* AtlasFlow only, *w/o AtlasFlow* PVA only, *Full* Ours), quantifies this. AtlasFlow alone reduces SeamL1/L2/Grad by 36.7%/36.0%/37.6% and DS by 13.7% relative to DiT360, with seven of the thirteen endpoints improving significantly; the complete method reduces SeamL1/L2/Grad by 30.8%/29.9%/37.7% and DS by 12.2% with five significant improvements. Across seeds SeamL1 spans 0.0151–0.0158 and DS 2.96–3.02 against DiT360's 0.0225 and 3.41 (Appendix L). Two controls in Appendix C tie the gain to the mechanism: the symmetric sampler has exactly zero half-turn discrepancy in velocity and latent trajectory where the released sampler retains a relative error of 0.34 (Table 7); on clean references the chart whose cut is farther from a region predicts it better at every noise level (Table 8), the assumption behind Eq. (5). The decoder choice alone moves the seam scores (Sec. 5.6), which is why every comparison holds the decoder fixed. The construction transfers to other frozen samplers (Appendix N). On frozen FLUX.1-dev, over the 20 prompts published with SphereDiff (Park et al., 2026) and reused by Hayakawa et al. (2026) (two seeds, n=40, no reference panoramas), it lowers DS from 71.21 to 3.12 and RingCut-P from 77.9 to 6.0 on identical weights, the level the panorama-tuned com-

Table 2: **Matched** 2×2 **component ablation on Matterport3D.** The rows form the complete 2×2 design under matched prompts, seeds, resolution, sampler, and decoder. *w/o PVA and AtlasFlow* is released DiT360; *w/o PVA* retains AtlasFlow only; *w/o AtlasFlow* retains PVA only; *Full* uses both. Gold, silver, and bronze denote the three best raw means within these four rows, not paired statistical significance.

Method	FID ↓	FID_{crop} ↓	FID_{equ} ↓	FAED ↓	BRISQUE ↓	NIQE ↓	Seam L1 ↓	Seam Grad ↓	VPFD ↓	VDRI ↓	VDAE ↓
w/o PVA and AtlasFlow	37.90	36.41	32.30	3.567	22.93	4.116	0.0225	1.625	5.748	6.305	0.0759
w/o PVA	37.49	35.82	**31.50**	3.471	22.96	4.129	**0.0143**	1.014	5.843	6.213	0.0749
w/o AtlasFlow	36.83	35.02	32.51	3.213	21.02	4.051	0.0239	1.574	**5.623**	6.306	0.0755
Full	**36.30**	**34.43**	31.75	**3.114**	**20.99**	**4.050**	0.0156	**1.012**	5.631	**6.150**	**0.0746**



Figure 4: **Leave-one-component-out qualitative ablation, both panels in the main configuration.** Pairs share prompt, seed, resolution, guidance, scheduler, and decoder and differ only in the named component. Cells: the ERP with the stored cut at the centre and, below it, a square 90° view (**blue** footprint). **(a)** At the cut camera (yaw -180°, dashed line): PVA alone leaves the dining room split at the cut, the complete method renders one continuous room (SeamL1 $0.0335 \rightarrow 0.0081$). **(b)** A Matterport3D bedroom at yaw -180°, pitch -45°: without PVA the bed is wedge-shaped with a curved foot edge, and the complete method recovers a straight one (VPFD at this view $10.3^\circ \rightarrow 2.5^\circ$). Both examples are metric-guided picks (Appendix O).

plete method reaches on the same prompts (2.97 and 5.6): the cut repair does not require panoramic pretraining.

5.3 Local geometry of rendered views

PVA acts on what a 90° camera sees. Figure 4(b) renders a Matterport3D bedroom 45° downward: without PVA the bed is wedge-shaped with a curved foot edge, and the complete method recovers a straight one, VPFD at this view falling from 10.3° to 2.5°. Over the cohort's 16 calibrated views per ERP (Table 2) the complete method improves VPFD, VDRI, and VDAE over DiT360 by 2.0%, 2.4%, and 1.7%, VDAE significant in the paired test and VPFD and VDRI shifting the same way. The same residual carries the reference-quality gains: added to the symmetric sampler it moves seven of the ten aggregate scores in the favorable direction (FID 37.49 to 36.30, FAED 3.47 to 3.11, BRISQUE 22.96 to 20.99).

5.4 Component ablation

The two components of Table 2 act on different endpoints. AtlasFlow supplies the cut correction (SeamL1 0.0225 to 0.0143) and leaves the aggregate scores where DiT360 has them; PVA supplies the reference-quality and viewport profile, with the teacher weight λ_{nat} setting the balance between the two (Appendix B).

5.5 Cross-dataset ODI-SR1200 evaluation

On ODI-SR1200 we use the same complete-method checkpoint and inference rule as on Matterport3D. Relative to the matched DiT360 baseline (FID 59.64, FAED 16.82), Ours reaches FID 55.02 and FAED 14.64. Across the thirteen pre-specified per-panorama endpoints, Ours records 12 significant improvements. SeamL1/L2/gradient decrease by 41.7%/49.1%/52.1%, StripL1 by 70.3%, VPFD by 6.5%, and VDRI/VDAE by 8.8%/5.4%. The training-free rule alone, with no learned parameters, already improves all four FIDs and FAED over the baseline (FID 56.00); the complete method adds a further 0.98 FID and improves the two no-reference scores (BRISQUE $24.59 \rightarrow 23.79$, NIQE $3.939 \rightarrow 3.920$); full tables are in Appendix D.

5.6 Boundary seam metrics miss a relocated seam

Two findings bound how seam scores should be read. First, rolling every generated image by $180°$, a lossless relabeling of longitude, drops the DS of released DiT360 from 16.73 to 2.97, comparable to the fused samplers, while a yaw-invariant scan is unchanged (Appendix M): a sampler that moves its seam away from the frame boundary passes DS. DS is also uncalibrated across datasets (real Matterport3D panoramas score 1.24, real ODI-SR panoramas 7.91; RingCut places both near one). Injecting seams of known strength into real panoramas quantifies this: DS stays near or below chance for cuts injected at interior longitudes, while the RingCut prominence read-out reaches AUC 0.92 and 0.996 at amplitudes of 5% and 20% of grayscale standard deviation on ODI-SR (Appendix G). Second, on identical latents the decoder choice alone changes DS from 9.48 to 3.70 and SeamL1 by 52% (Appendix F), so sampler comparisons must hold the decoder fixed, as all of ours do.

6 Conclusion

AtlasFlow fuses chart-indexed velocities before one shared update with exact latent half-turn and joint input–layout equivariance and no training; PVA adds an 18K-parameter distilled correction. On Matterport3D the complete method reduces SeamL1/L2/Grad by 30.8%/29.9%/37.7% at parity or better with DiT360 on every aggregate score but IS and CLIPScore, leads eight of ten metrics among twelve generators, and reaches ODI-SR1200 FID 55.02 versus 59.64. A yaw-invariant scan detects relocated synthetic seams where boundary metrics stay at chance.

Limitations. Neither continuous $SO(2)$ nor exact decoded-RGB equivariance is claimed (the planar VAE is only approximately cyclic). Two denoiser calls per step double sampler time (31.1 vs. 15.7 s per image, one H100). The native teacher sees pitch in $[-40°, 40°]$ only, so polar views are not directly supervised.



Figure 5: **Outputs of the complete method**, ERPs rolled so that the stored cut is the marked centre column: a Nuremberg square, a flooded library, a ski-resort terrace, a Christmas market, a historic interior, and a floating city; twelve more in Appendix O.

Ethics statement

Generated panoramas can depict plausible but inaccurate environments and should not be treated as measurements or documentary evidence of real locations. Pretrained-model biases and the usage conditions of the underlying models and datasets continue to apply. Improving geometric continuity does not certify the factual accuracy or safety of a generated scene.

Reproducibility statement

The method section defines the chart transforms, weighting rules, adapter objective, and order of inference operations. The appendix specifies the training and evaluation cohorts, the complete inference configuration, the PVA architecture and selected checkpoint, and the metric and paired-testing protocols. The main method is the symmetric theorem-bearing construction; the local-handoff variant is labeled wherever it appears. The figure sources include a vector generator for the pipeline, seven qualitative rendering scripts, and the stored-cut distance-profile computation that specify their arms, prompt identities, cameras, and selection rules. Their source images and the remaining decoder diagnostic without a plotting script are identified in the figure manifest. The pipeline's noise and velocity textures are schematic, not measured intermediate states. Raster reproduction can vary slightly with imaging-library versions; it is checked separately from source-image and camera provenance. The paired-endpoint tables are built from stored scorecards rather than transcribed.

References

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *Proceedings of the International Conference on Machine Learning*, 2023.

Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Black Forest Labs. FLUX.1 [dev]. `https://huggingface.co/black-forest-labs/FLUX.1-dev`, 2024.

Hakan Çapuk, Andrew Bond, Muhammed Burak Kızıl, Emir Göçen, Erkut Erdem, and Aykut Erdem. TanDiT: Tangent-plane diffusion transformer for high-quality 360-degree panorama generation. *arXiv preprint arXiv:2506.21681*, 2025.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *International Conference on 3D Vision*, pp. 667–676, 2017.

Anders Christensen, Nooshin Mojab, Khushman Patel, Karan Ahuja, Zeynep Akata, Ole Winther, Mar Gonzalez-Franco, and Andrea Colaco. Geometry fidelity for spherical images. In *European Conference on Computer Vision*, 2024. URL `https://arxiv.org/abs/2407.18207`.

Xin Deng, Hao Wang, Mai Xu, Yichen Guo, Yuhang Song, and Li Yang. LAU-Net: Latitude adaptive upscaling network for omnidirectional image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9189–9198, 2021.

Haoran Feng, Dizhe Zhang, Xiangtai Li, Bo Du, and Lu Qi. Dit360: High-fidelity panoramic image generation via hybrid training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23367–23377, 2026.

Mengyang Feng, Jinlin Liu, Miaomiao Cui, and Xuansong Xie. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. *arXiv preprint arXiv:2311.13141*, 2023.

Stanislav Frolov, Brian B. Moser, and Andreas Dengel. SpotDiffusion: A fast approach for seamless panorama generation over time. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2025.

Akio Hayakawa, Yusuke Mukuta, and Tatsuya Harada. Linear fusion MultiDiffusion for fast training-free spherical panorama generation. In *European Conference on Computer Vision (ECCV)*, 2026.

Or Hirschorn, Aaron Olender, Eli Alshan, Ianir Ideses, Lior Fritz, and Sagie Benaim. SpheRoPE: Zero-shot optimization-free 360 panorama generation with spherical RoPE. *arXiv preprint arXiv:2606.32033*, 2026.

HunyuanWorld Team. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025.

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

Jungwoon Kang, Jaehun Kim, Yiwon Yu, Hyungyum Jang, Sanghoon Lee, and Jongyoo Kim. SHERPA: Seam-aware harmonized ERP adaptation for open-domain 360-degree panorama generation. *arXiv preprint arXiv:2606.12213*, 2026.

Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. SyncDiffusion: Coherent montage via synchronized joint diffusions. In *Advances in Neural Information Processing Systems*, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL `https://arxiv.org/abs/2210.02747`.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 2012.

Anish Mittal, Rajiv Soundararajan, and Alan Conrad Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 2013.

Jinhong Ni, Chang-Bin Zhang, Qiang Zhang, and Jing Zhang. What makes for text to 360-degree panorama generation with stable diffusion? *arXiv preprint arXiv:2505.22129*, 2025.

Changgyoon Oh, Wonjune Cho, Yujeong Chae, Daehee Park, Lin Wang, and Kuk-Jin Yoon. BIPS: Bi-modal indoor panorama synthesis via residual depth-aided adversarial learning. In *European Conference on Computer Vision*, 2022.

Minho Park, Taewoong Kang, Jooyeol Yun, Sungwon Hwang, and Jaegul Choo. SphereDiff: Tuning-free 360-degree static and dynamic panorama generation via spherical latent representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8305–8313, 2026. doi: 10.1609/aaai.v40i10.37779.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Xiancheng Sun, Mai Xu, Shengxi Li, Senmao Ma, Xin Deng, Lai Jiang, and Gang Shen. Spherical manifold guided diffusion model for panoramic image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5824–5834, 2025.

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. MVDiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *Advances in Neural Information Processing Systems*, 2023.

Chaoyang Wang, Xiangtai Li, Lu Qi, Xiaofan Lin, Jinbin Bai, Qianyu Zhou, and Yunhai Tong. Conditional panoramic image generation via masked autoregressive modeling. arXiv preprint arXiv:2505.16862, 2025.

Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4933–4943, 2024a.

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

Yujia Wang, Yuyan Li, Jiuming Liu, Fang-Lue Zhang, Xinhu Zheng, and Neil A. Dodgson. RL-ScanIQA: Reinforcement-learned scanpaths for blind 360-degree image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37401–37412, 2026.

Chenxu Wu, Qingpeng Kong, Peiang Zhao, Wendi Yang, Wenxin Ma, Fenghe Tang, Zihang Jiang, and S. Kevin Zhou. Equivariant sampling for improving diffusion model-based image restoration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6827–6839, 2026a.

Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. PanoDiffusion: 360-degree panorama outpainting via diffusion. In *International Conference on Learning Representations*, 2024.

Ziyi Wu, Daniel Watson, Andrea Tagliasacchi, David J. Fleet, Marcus A. Brubaker, and Saurabh Saxena. 360Anything: Geometry-free lifting of images and videos to 360-degree. In *European Conference on Computer Vision*, 2026b.

Ziyang Xie. WorldGen: Generate any 3d scene in seconds. GitHub repository, `https://github.com/ZiYang-xie/WorldGen`, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems*, 2024a.

Shuai Yang, Jing Tan, Mengchen Zhang, Tong Wu, Yixuan Li, Gordon Wetzstein, Ziwei Liu, and Dahua Lin. LayerPano3D: Layered 3d panorama for hyper-immersive scene generation. arXiv preprint arXiv:2408.13252, 2024b.

Zhongqi Yang, Wenhang Ge, Yuqi Li, Jiaqi Chen, Haoyuan Li, Mengyin An, Fei Kang, Hua Xue, Baixin Xu, Yuyang Yin, Eric Li, Yang Liu, Yikai Wang, Hao-Xiang Guo, and Yahui Zhou. Matrix-3D: Omnidirectional explorable 3d world generation. arXiv preprint arXiv:2508.08086, 2025.

Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6347–6357, 2024.

Leo Zhang, Kianoosh Ashouritaklimi, Yee Whye Teh, and Rob Cornish. SymDiff: Equivariant diffusion via stochastic symmetrisation. In *International Conference on Learning Representations*, 2025.

Dian Zheng, Cheng Zhang, Xiao-Ming Wu, Cao Li, Chengfei Lv, Jian-Fang Hu, and Wei-Shi Zheng. Panorama generation from NFoV image done right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21610–21619, 2025.

Input: prompt c, shared initial packed latent z_T, scheduler times $\{t_k\}$, frozen panorama denoiser F_θ, optional frozen PVA A_ϕ, and chart offsets $G = \{0, \pi\}$.

1. Encode c once and reuse its text embeddings in every chart call.
2. For each scheduler time t_k and each $g \in G$, roll the same current latent to chart g, add the released periodic ghost columns, and evaluate F_θ with image IDs tied to chart-grid positions. Crop the ghost predictions and unroll the result to the canonical physical frame: $v_g = \rho_g^{-1} C F_\theta(P(\rho_g z_{t_k}), t_k, c, q)$.
3. At packed-column centers $u = (x+1/2)/W_p$ set $w_0 = \sin^2(\pi u)$ and $w_\pi = \cos^2(\pi u)$ (Eq. (6)). Fuse aligned velocities as $\bar{v} = w_0 v_0 + w_\pi v_\pi$. (The local-handoff variant instead uses $d = \min(u, 1-u)$, $a = \mathrm{clip}(d/r, 0, 1)$, $w_0 = 3a^2 - 2a^3$ with $r = 1/16$.)
4. Unpack $\bar{v}$ and z_{t_k} to the 16-channel VAE grid. For the complete method, apply the residual once after fusion, $\widetilde{V} = \bar{V} + A_\phi(\bar{V}, x_{t_k}, t_k)$, and repack it. For AtlasFlow-only, set $A_\phi = 0$.
5. Perform exactly one shared scheduler update $z_{t_{k+1}} = \mathrm{Step}(z_{t_k}, \mathrm{Pack}(\widetilde{V}), t_k)$.

After the final update, unpack z_0, append eight cyclic VAE-latent columns on each side, decode once, and crop the corresponding 64 RGB pixels per side.

Table 3: **Complete fixed-layout inference (symmetric default).** AtlasFlow changes the velocity used by the trajectory, whereas circular decoder context is a matched output policy applied to both baseline and candidate.

A Reproducibility and evaluation cohorts

Four data roles are kept apart: PVA optimization, sampler development, controlled in-domain evaluation, and cross-dataset evaluation. This distinction is important because the public Matterport3D-polished release contains one 10,359-image split rather than the original paper's undisclosed polished validation identifiers. We optimize PVA on that complete public release. The public MVDiffusion/PanFusion `test.npy` identifiers (Tang et al., 2023; Zhang et al., 2024) that are present in Matterport3D-polished (about 1k images) and their released captions form the Matterport3D evaluation cohort. ODI-SR1200 is the cross-dataset evaluation; layout, fusion-profile, and checkpoint comparisons are also reported on it.

All matched generator comparisons share prompt text, per-image seed, resolution, number of denoising steps, guidance, scheduler, base checkpoint, image-token convention, padding refresh, and decoder policy. Their manifests store the ordered prompt, seed, reference identifier when available, and every sampler switch. No row is silently dropped when a per-image diagnostic fails; metrics with invalid views report the common-valid coverage explicitly. Some archived DiT360 manifests retain one-time duplicate-token padding rather than per-step refresh. We checked the corresponding saved refresh controls: all Matterport3D-cohort and all 1,000 OOD1000 PNGs are byte-identical to their one-time-padding counterparts. Those baseline artifacts therefore represent the same generated images despite the differing manifest flag.

Complete inference procedure. Table 3 states the implementation-level order. It is crucial that the two chart predictions update one shared latent: we never generate two decoded images and cross-fade them. At 2048×1024, the VAE latent is $16\times128\times256$ and FLUX packing gives a 64×128 token grid with 64 features per token. The local-handoff variant's $r = 1/16$ band therefore spans eight packed columns, or 128 RGB pixels, on each side of the stored cut.

B PVA configuration and controlled selection

All controlled DiT360 comparisons fix prompts, per-image seeds, resolution, inference steps, guidance, base checkpoint, image-token convention, scheduler, and decoder policy in a machine-readable manifest.

Architecture and optimization. The reported Perspective Velocity Adapter (PVA) has 18,128 trainable parameters. Its input is the concatenation of the 16-channel frozen velocity, 16-channel noisy VAE latent, and one normalized timestep plane. A 1×1 convolution projects to 64 channels; three depthwise-separable residual blocks use 3×3 kernels with spatial dilations $1, 2, 4$; a zero-initialized 1×1 projection returns a 16-channel residual. Width padding is circular and latitude

Layer	Shape / operation	Params.
Input projection	$33 \to 64$, 1×1	2,176
Periodic residual block $\times3$	GN, DW-3×3, PW-1×1	$3\times4,928$
Output normalization	GroupNorm(8, 64)	128
Output projection	$64 \to 16$, 1×1	1,040
Total		**18,128**

Table 4: **Exact PVA parameter count.** DW denotes a 64-group depthwise convolution and PW a pointwise convolution. The final projection is zero-initialized, so the adapter initially leaves the frozen velocity unchanged.

padding replicates the boundary. The selected native-view configuration has no longitude coordinate conditioning and no seam-local output gate. We freeze the original FLUX parameters and the released 74.7M DiT360 LoRA parameters and optimize only PVA with learning rate 10^{-4}, batch size one, gradient accumulation four, and mixed precision. Table 4 gives the exact parameter count layer by layer.

With stride one, the convolutional path spans $1 + 2(1 + 2 + 4) = 15$ VAE-latent cells on each axis, approximately 120×120 RGB pixels. This is the spatial footprint of the convolution kernels, not a strict dependency bound on the complete adapter: GroupNorm also uses spatial statistics, and the input velocity already contains global denoiser context. PVA supplies a compact residual correction; global scene layout and prompt semantics remain primarily the responsibility of the frozen denoiser.

The selected configuration uses native-teacher weight 0.10, random seed 281401, and the checkpoint after optimizer step 5,178. The auxiliary weight is linearly warmed for 1,000 steps. Every minibatch samples one 64×64 tangent latent with uniform yaw, uniform pitch in $[-40^\circ, 40^\circ]$, and a fixed 90° field of view. The same frozen DiT360 transformer is queried on that noisy crop with ordinary grid image coordinates, no ERP padding, and all panorama adapters disabled. Its velocity is the stopped target for the student velocity projected through the same tangent grid. The prompt, timestep, and noisy tangent latent are shared. No decoded RGB or external teacher is used by this loss, and the extra teacher call is discarded at inference.

Training-camera convention. The checkpoint uses the original tangent-grid implementation: with forward vector f, horizontal right vector r, and image-row coordinates increasing downward, it sets $u = r \times f$ and negates the row coordinate before forming the ray. Relative to an upright pinhole camera this vertically reflects the sampled square. The noisy latent and student velocity undergo the same reflection, so their correspondence is preserved, but the frozen teacher is not assumed to be reflection-equivariant. Thus the reported checkpoint tests a shared-grid consistency regularizer, not a separately validated upright-view teacher. This convention is retained for reproducing the results; training an otherwise matched adapter with $u = f \times r$ is an untested control.

The teacher target is a regularizer, not an exact change-of-variables target for a native-image diffusion process. Even with an upright camera, bilinear sampling Π_i maps white latent noise to noise with covariance $\Pi_i\Pi_i^\top$, generally not the identity, and a projected ERP latent need not equal the VAE encoding of the corresponding RGB perspective view. The supervised ERP flow loss remains the anchor; the auxiliary square-view prediction supplies an empirical consistency signal under this mismatch.

Current symmetric component design. Table 5 completes the main configuration's four-arm comparison on all thirteen paired diagnostics. Adding PVA to symmetric AtlasFlow improves seven standard metrics by raw value and reduces VPFD by 3.63% in a direct paired test (mean reduction 0.212°, 95% interval $[0.116^\circ, 0.310^\circ]$). These contrasts isolate what the adapter changes in the selected symmetric sampler rather than transferring evidence from the local variant.

Configuration comparison. Table 6 completes the component and teacher-weight ablation. Its first panel reports the nine retained DiT360 metrics for the two-component factorial at $\lambda_{\mathrm{nat}} = 0.10$ and the corresponding 0.05 control; its fusion rows use local handoff rather than the symmetric default. The remaining panels hold that local sampler fixed to isolate the teacher weight on all fifteen pre-specified paired endpoints. Both complete stacks improve eight of the nine retained

Endpoint	DiT360	PVA	AtlasFlow	Full
CS ↑	0.3436	0.3435	0.3435	0.3433
SeamL1 ↓	0.02252	0.02391‡	0.01426†	0.01559†
SeamL2 ↓	0.00293	0.00315‡	0.00187†	0.00205†
SeamGrad ↓	1.6247	1.5742†	1.0137†	1.0123†
StripL1 ↓	0.01678	0.01678	0.02075‡	0.02086‡
DS ↓	3.4115	3.4768‡	2.9445†	2.9946†
RingCut ↓	1.2573	1.2557	1.2557	1.2554
PanoSim ↑	0.2606	0.2596‡	0.2628†	0.2612
PanoProb ↑	0.5219	0.5200‡	0.5260†	0.5231
PanoMargin ↑	0.00880	0.00808‡	0.01047†	0.00932
VPFD ↓	5.7479	5.6232†	5.8431	5.6309
VDRI ↓	6.3047	6.3061	6.2135	6.1503
VDAE ↓	0.07587	0.07548	0.07492	0.07458†

Table 5: **Complete paired diagnostics for the symmetric component design.** Absolute means weight each panorama equally; †/‡ denote improvement/degradation against DiT360. Tests use 50,000 paired panorama-cluster bootstraps and endpointwise 95% intervals, without familywise correction. Depth comparisons pool common-valid overlap pairs, so their paired percentages need not equal percentages of the independently averaged absolute columns. The three panorama-likeness scores are correlated.

Method	FID ↓	$\mathrm{FID}_{\mathrm{crop}}$ ↓	$\mathrm{FID}_{\mathrm{pole}}$ ↓	$\mathrm{FID}_{\mathrm{equ}}$ ↓	FAED ↓	IS ↑	CS ↑	BRISQUE ↓	NIQE ↓
DiT360	37.90	36.41	62.30	32.30	3.567	5.359	**0.3436**	22.93	4.116
PVA, $\lambda_{\mathrm{nat}} = 0.05$	37.35	35.84	62.37	32.26	3.356	**5.443**	0.3434	22.19	4.091
PVA, $\lambda_{\mathrm{nat}} = 0.10$	36.83	35.02	62.11	32.51	3.213	5.401	0.3435	21.02	4.051
Local handoff	37.93	36.31	**61.63**	32.13	3.500	5.357	0.3433	22.81	4.114
+ PVA, $\lambda_{\mathrm{nat}} = 0.05$	37.37	35.67	62.14	32.13	3.291	5.432	0.3434	21.94	4.088
+ **PVA,** $\lambda_{\mathrm{nat}} = 0.10$	**36.63**	**34.82**	62.25	**32.08**	**3.164**	5.396	0.3430	**20.80**	**4.044**

Method	SeamL1 ↓	SeamL2 ↓	SeamGrad ↓	StripL1 ↓	RingCut ↓	DS ↓	PanoSim ↑	PanoProb ↑	PanoMargin ↑
Local handoff	**0.01392**	**0.001711**	**0.98826**	**0.01931**	**1.25735**	**2.9451**	**0.26095**	**0.52304**	**0.009287**
+ PVA, $\lambda_{\mathrm{nat}} = 0.05$	0.01470	0.001846	0.99221	0.01974	1.25814	2.9971	0.26075	0.52210	0.008910
+ **PVA,** $\lambda_{\mathrm{nat}} = 0.10$	0.01535	0.001894	0.99240	0.01965	1.25768	3.0112	0.25954	0.52028	0.008173

Method	VPFD ↓	VDRI ↓	VDAE ↓	Direct W/T/L
Local handoff	5.7232	6.1695	0.074831	–
+ PVA, $\lambda_{\mathrm{nat}} = 0.05$	5.6938	6.1664	0.074744	0/10/5
+ **PVA,** $\lambda_{\mathrm{nat}} = 0.10$	**5.5841**	**6.1097**	**0.074438**	1/7/7

Table 6: **PVA component and teacher-weight ablation on the Matterport3D evaluation cohort.** Metric entries are absolute means. In the first panel, bold marks the best raw value across the displayed component and weight rows; in the other panels, it marks the best value in the three-row fixed-local sweep. Each adapter row uses one checkpoint consistently across every metric family. Direct W/T/L compares that row with local handoff alone on the fifteen pre-specified paired endpoints using 95% panorama-level bootstrap intervals; the verdict counts were registered over the original fifteen endpoints and are unchanged. The sampler in this table is the local-handoff variant, with which the teacher-weight selection was run; the symmetric default inherits the selected weight (its own factorial is Table 2).

metrics over DiT360. We select $\lambda_{\mathrm{nat}} = 0.10$ because it gives the strongest FID, FAED, BRISQUE, and NIQE profile and a significant VPFD reduction relative to AtlasFlow. The 0.05 setting retains more of AtlasFlow's continuity scores, while the 0.10 setting provides the stronger local-camera result. Raw dataset-level shifts and paired endpoint significance answer different questions and are reported separately.

This analysis uses the Matterport3D evaluation cohort for controlled configuration selection and component attribution rather than as an independent cross-dataset estimate of PVA. The selected teacher weight, seed, and checkpoint are frozen as one configuration and are never switched across

Sampler	Flow $\mathcal{D}_{\mathrm{cut}} \downarrow$	Final $\mathcal{D}_{\mathrm{cut}} \downarrow$	Calls
Released	0.12274	0.34469	1
Local handoff	0.10165	0.31199	2
Symmetric (default)	**0**	**0**	2

Table 7: **Actual-model cut equivariance on eight fixed prompts.** Both discrepancy columns apply the relative half-turn criterion from Eq. (2): "Flow" to the predicted sigma-1 vector field before an update and "Final" to the matched eight-step packed-latent trajectory. "Calls" is transformer calls per step. Local handoff denotes the prior-preserving variant. The zeros for the simultaneous symmetric construction, our default, are exact by construction rather than rounded measurements; temporal alternation is included in the additional controls below.

σ	Opposite at cut 0	Chart 0 at cut π
0.9	2.19 [0.96, 3.34]	1.51 [0.09, 2.84]
0.5	2.95 [2.74, 3.16]	2.79 [2.60, 2.97]
0.1	10.29 [9.94, 10.62]	10.24 [9.84, 10.63]

Table 8: **Cut reliability on the first 60 public-test references.** Entries are mean percent reductions in squared target-flow error with 95% panorama-bootstrap intervals. The target is $\epsilon - z_{\mathrm{clean}}$ at a known noisy latent. Both pre-specified directions support downweighting a chart near its own cut; this is a mechanism test, not an image-quality score.

metric families. The paired evidence attributes a significant VPFD reduction to PVA; the seven-of-nine favorable shift over local handoff is separately interpreted as a raw reference/quality profile. The complete PVA–AtlasFlow configuration is evaluated on ODI-SR1200 without optimizing adapter weights on that dataset; additional inference and checkpoint comparisons are reported separately.

C Mechanism and implementation analysis

The four controls summarized in Sec. 5.2 are reported here in full. In the tables of this section, *Symmetric* denotes the default symmetric PoU rule and *Local handoff* the prior-preserving local-handoff variant.

On eight fixed prompts, the local-handoff variant reduces the 180° relative flow error from 0.12274 to 0.10165 and the matched eight-step packed-latent trajectory error from 0.34469 to 0.31199. The simultaneous symmetric construction, our default rule, has zero error in both tests, as predicted by Proposition 1. Table 7 reports both discrepancies and the call counts.

A target-flow test on clean reference panoramas independently checks the weighting assumption (Table 8). Across three noise levels and both pre-specified directions, the chart whose artificial cut is farther from the evaluated region has lower squared target-flow error. This supports downweighting a chart near its own cut; it is a mechanism test rather than an image-quality score.

Seed-disjoint replication. We also repeat matched generation with disjoint per-prompt seeds and analyze the two cohorts separately rather than treating repeated prompts as independent. The stored-cut reductions reproduce on both the development and Matterport3D cohorts. The complete seed-disjoint replication is reported in Appendix L.

D ODI-SR1200 aggregate and paired endpoint audit

ODI-SR1200 uses the 1,200 high-resolution panoramas from the official ODI-SR `training/HR` split; this is the source dataset's split name, and neither generator is optimized on these images. Following the public SpheRoPE protocol, Qwen3-VL-32B-Instruct descriptions are summarized by Qwen3-8B to fewer than 50 words. The matched DiT360 and frozen complete-method arms share ordered prompts, seeds 0–1199, 2048×1024 resolution, 28 denoising steps, guidance 3.0,

Table 9: **ODI-SR1200 comparison on the nine surviving DiT360 metrics.** Gold, silver, and bronze denote raw first, second, and third ranks among the displayed systems and do not imply statistical significance. All methods use the same 1,200 prompt/reference identities and metric implementation.

Method	FID ↓	$\mathrm{FID}_{\mathrm{crop}}$ ↓	$\mathrm{FID}_{\mathrm{pole}}$ ↓	$\mathrm{FID}_{\mathrm{equ}}$ ↓	FAED ↓	IS ↑	CS ↑	BRISQUE ↓	NIQE ↓
DiT360	59.64	55.18	88.2	49.27	16.82	**19.40**	**0.318391**	23.79	**3.920**
SpheRoPE	74.70	70.38	154.4	66.33	30.34	16.60	0.315256	27.22	3.999
Ours	**55.02**	**50.41**	**80.4**	**46.16**	**14.64**	18.32	0.314940	**23.35**	3.935

Endpoint	Better	DiT360	Ours	Verdict
CLIPScore	↑	0.318391	0.314940	Loss
SeamL1	↓	0.053438	0.031131	Win
SeamL2	↓	0.009807	0.004991	Win
SeamGrad	↓	2.0994	1.0052	Win
StripL1	↓	0.063454	0.018872	Win
PanoSim	↑	0.194530	0.207141	Win
PanoProb	↑	0.403431	0.423363	Win
PanoMargin	↑	−0.040378	−0.031850	Win
RingCut	↓	1.1909	1.1701	Win
DS	↓	3.6232	3.4206	Win
VPFD	↓	7.3110	6.8336	Win
VDRI	↓	5.6381	5.1410	Win
VDAE	↓	0.071206	0.067351	Win

Table 10: **Complete ODI-SR1200 paired endpoint audit.** Verdicts use 50,000 panorama-cluster bootstrap replicates and a 95% interval for the paired difference. Ours denotes the frozen PVA–AtlasFlow stack.

the released DiT360 checkpoint, image-token convention, per-step padding, and circular decoding. Official FLUX.1 SpheRoPE outputs use the same prompt/reference identities and metric code.

Table 9 compares the frozen complete method with DiT360 and the official SpheRoPE configuration on the same 1,200 prompts and references.

Table 10 gives the complete paired verdict (12 significant wins among the 13 endpoints). The endpoint set and bootstrap rule are fixed before inspecting the ODI-SR1200 outcomes. Dataset-level Fréchet distances are excluded from this count because they do not admit a per-panorama paired bootstrap under the reported implementation.

The depth comparison uses the 97.89% of overlapping-view pairs on which both arms produce valid depth predictions (every panorama contributes at least one valid pair); other paired endpoint families have full coverage. This coverage rule is fixed symmetrically and does not replace failed values with favorable constants.

E METRIC DEFINITIONS AND STATISTICAL PROTOCOL

The main reference/quality table reports nine metrics of the released DiT360 evaluator and adds KID (Binkowski et al., 2018), for ten metrics in total. KID is the unbiased squared MMD estimate on the same 2,048-dimensional Inception features as FID, using $k(x, y) = (x^\top y/2048+1)^3$, averaged over 100 independently drawn subsets of 1,000 images per arm and displayed multiplied by 10^3. Subsets are sampled without replacement within each draw; their variation is not a confidence interval over generation seeds. Historical ablation and seed tables without KID are explicitly nine-metric subsets. The IS column uses the scene-classification version of the score: a frozen Places365 ResNet-50, resize to 256 followed by a 224×224 center crop, and $\exp(\mathbb{E}_x \mathrm{KL}(p(y|x)\|p(y)))$, averaged over ten image splits. It is not an ImageNet Inception-v3 classifier score; all compared methods use the same backend. The released evaluator's "FIDclip" is Inception-feature FID after removing the top and bottom 12.5% of the ERP; we denote it by $\mathrm{FID}_{\mathrm{crop}}$. $\mathrm{FID}_{\mathrm{pole}}$ uses the concatenated top and bottom 12.5% bands, while $\mathrm{FID}_{\mathrm{equ}}$ uses the central 25% equatorial band. These correspond to $|\mathrm{latitude}| > 67.5^\circ$ and $|\mathrm{latitude}| < 22.5^\circ$, respectively; the cropped variant retains $|\mathrm{latitude}| < 67.5^\circ$. They are ERP bands, not pinhole projections, and do not directly measure viewport geometry. FID itself is the Fréchet distance between Gaussians fitted to Inception-v3 pool features of the generated and

reference sets, so it falls as the two feature distributions approach each other; FAED (Oh et al., 2022) is the same distance in the encoder features of the BIPS panorama autoencoder, which is trained on ERPs and therefore responds to layout and boundary errors that Inception features ignore. IS rewards images that the Places365 classifier assigns confidently to one scene class while the classes are spread across the set. BRISQUE (Mittal et al., 2012) fits natural-scene statistics of locally normalized luminance and scores their departure from natural images with a learned regressor; NIQE (Mittal et al., 2013) measures the distance of the same statistics from a natural-image model without any learned quality regression. Both are no-reference scores of the decoded ERP, lower is better, and both respond to resolution and sharpness (Appendix H), which is why they are read together with the reference-based columns.

For a normalized RGB panorama $I \in [0,1]^{H\times W\times 3}$, the direct stored-cut errors are

$$\mathrm{SeamL1}(I) = \frac{1}{3H}\sum_{y,c} \left| I_{y,0,c} - I_{y,W-1,c} \right|, \tag{11}$$

$$\mathrm{SeamL2}(I) = \frac{1}{3H}\sum_{y,c} \left(I_{y,0,c} - I_{y,W-1,c} \right)^2. \tag{12}$$

SeamGrad divides SeamL1 by the mean absolute RGB difference between all adjacent horizontal columns, with a numerical ϵ in the denominator. DS applies a horizontal Scharr operator to the six-column wrap neighborhood and normalizes the two seam-adjacent responses by their outward context (Christensen et al., 2024). StripL1 compares the mean RGB colors of the left and right 16-column strips. RingCut computes the same context-normalized two-sided Scharr response at every cyclic longitude, applies $\log(1+x)$, averages over rows, and reports the mean of the strongest one percent of longitude cuts. We use the fixed `mean_top1pct` variant. These quantities need not vanish on a continuous ERP because the first and last pixel centers correspond to adjacent rather than identical rays.

Panorama-likeness uses CLIP ViT-B/32 with its standard image processor and two fixed texts: "a 360 degree equirectangular panorama image" and "a normal perspective photograph." Let their normalized image–text cosine scores be s_{pano} and s_{persp}. PanoSim is s_{pano}, PanoMargin is their difference, and PanoProb is $\mathrm{sigmoid}(10\,\mathrm{PanoMargin})$. The latter is a two-label softmax score, not a calibrated probability that a panorama is valid. The three read-outs are correlated and are not three independent kinds of geometric evidence. CS likewise reports raw normalized image–prompt cosine, without the optional scaling used in other CLIPScore conventions.

For local-view evaluation, we render 16 calibrated 90° pinhole views per ERP. VPFD compares frozen Perspective-Fields up-vector and latitude predictions with the analytic fields of the known camera (Jin et al., 2023). VDRI and VDAE compare Depth Anything V2 predictions from overlapping views on matched spherical rays using rank consistency and robust affine-aligned residuals, respectively (Yang et al., 2024a). OmniFID compares tangent/cubemap feature distributions (Christensen et al., 2024). Per-image metrics use paired bootstrap resampling over panoramas; a win or loss requires the 95% interval of the paired difference to exclude zero. Dataset-level Fréchet distances are ranked by raw value and are not assigned paired significance. Intervals are endpoint-wise 95% intervals without a family-wise multiplicity correction; win counts summarize this declared diagnostic panel, not independent confirmations of a general superiority claim.

More explicitly, VPFD is half the mean up-vector angular error plus half the mean latitude absolute error, both in degrees. For matched rays in a pair of overlapping views, relative camera disparity is multiplied by the ray's forward cosine to obtain inverse radial depth. We remove samples outside either prediction's 1st–99th percentile range. If ρ is the Spearman correlation of the remaining depths, VDRI is $50(1-\rho)$; zero indicates agreement of ordering, not correct metric depth. For VDAE, let $e(a,b)$ be the median absolute residual of an affine fit from a to b, divided by b's 10th–90th percentile range. We report $\frac{1}{2}[e(a,b)+e(b,a)]$. The affine fit uses five iteratively reweighted least-squares steps with Huber weights (threshold 1.345 times a MAD-based scale) and does not constrain the slope's sign. Thus VDAE can miss a depth ordering reversal; VDRI is retained as a complementary check. Both assume gravity-aligned panoramas and a shared optical center, and neither proves that the scene is globally realizable.

The component table averages each panorama's valid depth-overlap scores and then averages panoramas. The paired depth tests instead retain only overlap pairs valid in both arms, pool these

matched pairs for the point estimate, and resample entire panoramas for the confidence interval. Missing overlap pairs therefore produce small differences between absolute and paired-table means (for example, Full VDRI is 6.150 in the component table and 6.158 in the paired test), without implying a different checkpoint. The paired Full–DiT360 comparison retains 99.80% of scheduled overlap pairs.

F What pixel-level seam metrics measure

Every seam quantity in this paper (SeamL1, SeamL2, SeamGrad, DS, and RingCut; Appendix E) is read from decoded RGB, but the sampler only ever produces latents. Between the two sits the VAE decoder, which includes convolutions and attention. Circular latent context presents the wrap junction as an interior region and lets the decoder use both sides; it can lower a pixel step without changing the input latent. The two DiT360 rows of the decoder control isolate this effect: they decode bitwise-identical latents (same seeds, no chart fusion, released duplicate-token padding) and differ only in the decoder policy. Circular-8 decoding alone lowers SeamL1 from 0.0648 to 0.0314 (-52%, lower on all 60 images), SeamGrad from 2.58 to 1.25, and DS from 9.48 to 3.70 (-61%), while the yaw-invariant RingCut moves by 1.2% and CLIP prompt alignment is unchanged (0.3546 vs. 0.3547, paired tie). On the 24-prompt development set the same switch gives SeamL1 0.0778 to 0.0410 (-47%). The sampler's own contribution is the difference between the two Circular-8 rows (SeamL1 0.0314 to 0.0254).

Latent level. A video VAE makes the point sharper because each latent column decodes to 32 pixel columns. In a transfer experiment on the frozen LTX-Video prior (HaCohen et al., 2025) (no panorama training; 84 prompts $\times$ five seeds, 768×384, 25 frames; median over videos), the first/last-column pixel discontinuity of the raw prior is $6.80\times$ the interior level and drops to $0.97\times$ when the same latents are decoded with four columns of circular context, while the latent discontinuity is $1.53\times$ in both cases, bitwise identical. Per-step circular padding, which acts inside the sampler, is what lowers the latent figure ($1.03\times$). A pixel-level seam number therefore certifies the decoder as much as the sampler.

Consequences for evaluation. (i) *Attribution.* In the OOD60 DiT360 control above, circular decoding alone reduces SeamL1 by 52%; this does not specify the fraction of a different method's gain attributable to its decoder. A seam number is not comparable across methods or papers unless the decoder policy is stated and matched, which is why the main DiT360–AtlasFlow comparisons fix Circular-8 for both arms and credit AtlasFlow only with the matched-row difference (matched decoder control, the decoder control). Circular decoding itself is prior practice (Wu et al., 2026b), not a contribution. (ii) *Locality.* The added context is eight latent columns per side, corresponding to 64 pixels at the $8\times$ VAE; this is a padding width, not the decoder's receptive-field bound. Decoding cannot make the two sides of the input latent agree, and any further latent-space processing consumes the unchanged latent. Figure 6 illustrates the RGB change under the controlled same-latent decode. (iii) *Scope.* For a viewer of one decoded image the distinction is immaterial: a hidden step is hidden. It matters when the number is used to rank samplers, which is what seam tables do. The input-latent statistics are unchanged by decoding; RGB RingCut changes less in this control but is not decoder-invariant.

G What boundary seam metrics can and cannot detect

The stored-cut metrics (SeamL1, SeamL2, SeamGrad, DS) read the panorama at one longitude, the frame boundary. To measure what they detect, we inject a seam of known strength into real panoramas and ask whether a metric can tell the injected image from the clean one. For grayscale intensities $G \in [0, 1]$ and a cut column c, the injection is a cyclic sawtooth,

$$G_\alpha(y, j) = \mathrm{clip}_{[0,1]}\left[G(y, j) + \alpha\ \mathrm{std}(G)\left(\frac{(j - c) \bmod W}{W} - \frac{1}{2}\right)\right]. \tag{13}$$

Its jump is at c and its compensating ramp spans the full width; c is either the frame boundary or a longitude drawn uniformly from the interior $[W/8,\ 7W/8)$. We then compute the area under the ROC curve (AUC) of the metric for separating injected from clean images, pooled over all clean

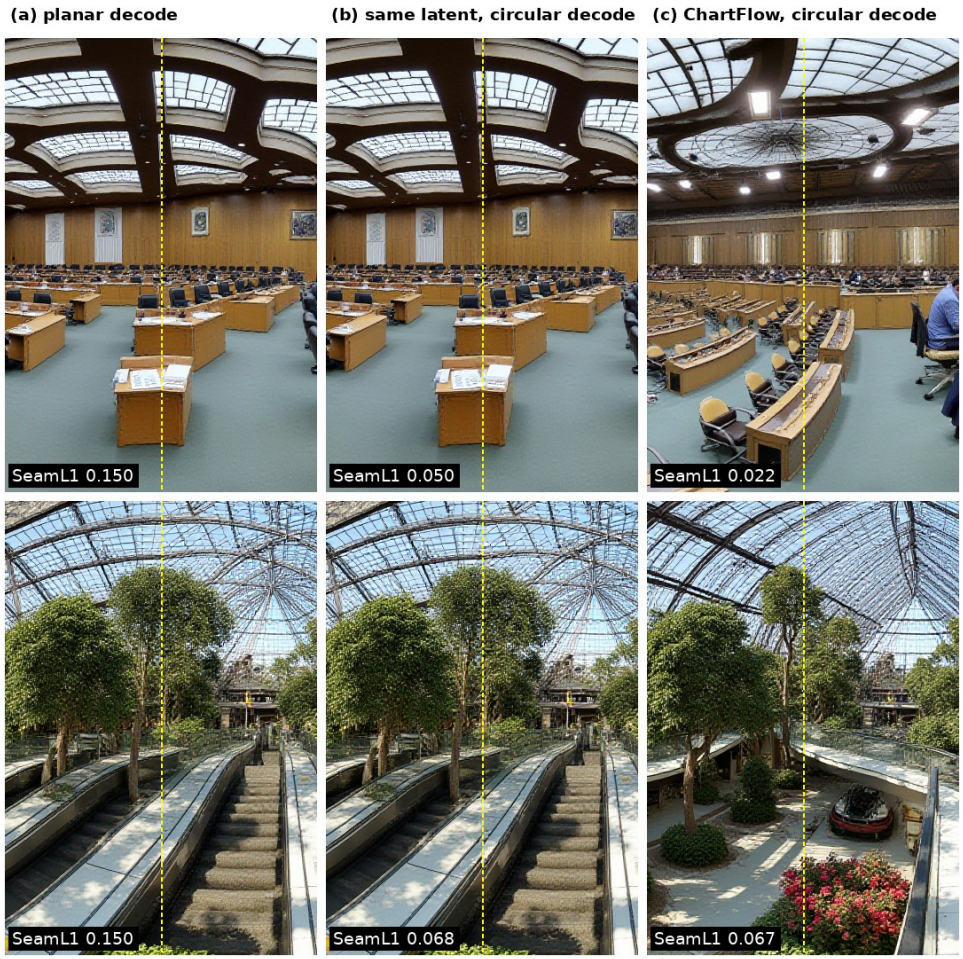

Figure 6: **The decoder hides the pixel step; it does not reconcile the latent.** Wrap junction (last and first 176 columns, equatorial band) of two OOD60 panoramas. (a) The released DiT360 latent decoded without circular context shows a hard line at the stored cut. (b) The bitwise-identical latent decoded with eight columns of circular context: the hard line is attenuated. (c) The AtlasFlow sample for the same prompt and seed under the same decoder, a different image because the sampler differs.

and injected images rather than paired per image, so that 0.5 is chance. Table 11 reports DS and the yaw-invariant RingCut scan on 100 Matterport3D and 100 ODI-SR panoramas at 2048×1024. This localization experiment uses a coherent scan (median rather than mean over rows); it is distinct from the `mean_top1pct` endpoint above. Both use a 10^{-3} stabilizer in the Scharr context denominator and average the largest $\lceil 0.01W \rceil$ longitude responses after $\log(1 + x)$.

Two conclusions follow. First, DS does not inspect a remote cut directly: on Matterport3D its AUC stays between 0.50 and 0.55 at every strength, and on ODI-SR it moves below chance. (The below-chance values are partly a property of our injection model, whose ramp raises the context term in the DS normalizer; these results show a failure to detect off-boundary seams in this injection family, not universal anti-correlation.) A boundary seam is detected instead: on Matterport3D with AUC 1.00 from $\alpha = 0.01$ upward, and on ODI-SR more weakly at small amplitudes (0.65 at $\alpha = 0.01$, rising to 0.99 at $\alpha = 0.2$). This is the quantitative form of the roll observation in the main text: a sampler that moves its seam away from the frame boundary passes DS unchanged. Second, RingCut is far better at locating a seam than at scoring it: for $\alpha \geq 0.1$ its argmax recovers the injected longitude in 83–100% of images, while the AUC of its numerical value reaches only 0.67–0.88. Its value is dominated by image-to-image differences in natural edge content rather than by the seam itself.

Table 11: **Detection sensitivity of seam metrics to an injected seam of relative strength α.** AUC of separating seam-injected from clean real panoramas (0.5 = chance) for the frame-boundary DS score and the yaw-invariant RingCut score, when the seam is injected at the frame boundary (b) or at a uniformly random longitude (r); "loc." is the fraction of injected seams whose longitude is recovered by the RingCut argmax. $n = 100$ real panoramas per dataset at 2048×1024.

Dataset, α	DS (b)	DS (r)	RingCut (b)	RingCut (r)	loc.
Matterport3D, 0.005	0.997	0.503	0.491	0.526	0.12
, 0.01	1.000	0.504	0.495	0.539	0.25
, 0.02	1.000	0.505	0.518	0.565	0.49
, 0.05	1.000	0.510	0.605	0.647	0.86
, 0.1	1.000	0.519	0.695	0.733	0.94
, 0.2	1.000	0.530	0.787	0.813	0.99
, 0.4	1.000	0.551	0.861	0.882	1.00
ODI-SR, 0.005	0.591	0.487	0.506	0.501	0.13
, 0.01	0.649	0.481	0.508	0.505	0.20
, 0.02	0.748	0.465	0.519	0.520	0.52
, 0.05	0.911	0.422	0.569	0.572	0.90
, 0.1	0.972	0.362	0.638	0.638	1.00
, 0.2	0.988	0.280	0.712	0.712	1.00
, 0.4	0.993	0.201	0.786	0.781	1.00

Table 12: **RingCut read-outs under seam injection at a random longitude ($n = 60$ per cohort, 1024×512).** AUC against the clean panoramas. top1 is the reported mean of the strongest 1% of cuts; prom is the peak prominence $(\max - \text{median})/\text{MAD}$ of the same per-longitude profile; ms denotes a multi-scale profile. The prominence read-out leads on both cohorts once the injected step reaches 5% of image contrast; below that the two read-outs are indistinguishable on Matterport3D, whose indoor scenes carry many genuine full-height edges.

α	DS	RingCut top1	RingCut-P (prom)	ms top1	ms prom
ODI-SR					
0.01	0.474	0.510	**0.556**	0.533	0.539
0.02	0.459	0.552	**0.694**	0.580	0.646
0.05	0.430	0.671	**0.925**	0.740	0.912
0.1	0.414	0.787	**0.984**	0.855	0.974
0.2	0.373	0.884	**0.996**	0.930	0.991
Matterport3D					
0.01	0.478	0.508	0.500	0.499	0.500
0.02	0.463	0.522	0.519	0.506	0.506
0.05	0.427	0.587	**0.659**	0.558	0.606
0.1	0.385	0.673	**0.813**	0.657	0.759
0.2	0.325	0.759	**0.952**	0.767	0.919

A prominence read-out (RingCut-P). If the location is right but the value is not, the profile should be normalized within each image. Let $p(\phi)$ be RingCut's per-longitude profile (the context-normalized two-sided Scharr response, median over rows, after $\log(1 + x)$). Instead of the mean of the strongest one percent of longitudes, we report the peak prominence

$$\text{RingCut-P}(I) = \frac{\max_\phi p(\phi) - m}{\max(\text{MAD}_\phi\, p(\phi), 10^{-3} \max(m, 10^{-6}))}, \qquad m = \text{median}_\phi\, p(\phi), \tag{14}$$

where MAD is the median absolute deviation of the same profile; its floor prevents unstable division for nearly constant profiles. The statistic emphasizes isolated peaks rather than a uniformly high edge background. Natural structures can also produce such peaks, so it is not a semantic classifier of seam artifacts. Table 12 compares the read-outs on 60 ODI-SR panoramas at 1024×512: at $\alpha = 0.05$ the prominence read-out raises the detection AUC from 0.67 to 0.92, and it reaches 0.996 at $\alpha = 0.2$, whereas a multi-scale profile helps the mean read-out only modestly and does not improve prominence. RingCut-P requires no learned model, no re-projection and no GPU; it is a one-line change of the read-out. The same ordering holds on the Matterport3D cohort above

Severity family	VPFD-90	VDRI-90	VDAE-90
Longitude warp	0.50	0.90	0.90
Vertical wave	0.90	0.90	1.00
Seam mismatch	0.60	0.80	1.00
Gaussian blur control	0.90	1.00	1.00

Table 13: **Controlled diagnostic validation.** Spearman correlation between four corruption levels plus clean and each lower-is-better tail score on eight fixed real panoramas. A valid 45° yaw changes VPFD-90, VDRI-90, and VDAE-90 by -0.663, $+1.126$, and -0.0010, respectively.

$\alpha = 0.05$; below it the two read-outs tie there, which is the expected cost of discounting the many genuine full-height edges of indoor scenes.

It is a paired statistic, not a cross-method score. The injected-seam test compares an image with itself, and the improvement above is a statement about that setting only. Computed across ten generators — the eight external systems PanFusion, MVDiffusion, WorldGen, Matrix-3D, LayerPano3D, SMGD, PAR and HunyuanWorld, plus UniPano (Ni et al., 2025) and AtlasFlow without PVA — and the real panoramas, the read-out does not rank them by seam quality. Its per-image denominator barely moves between systems (median MAD 0.0335–0.0385), so the cohort ordering is almost entirely the numerator: $r = 0.993$ between a system's median peak height and its median prominence. The statistic therefore rewards outputs that contain no strong vertical structure anywhere, whatever the reason. The lowest value (6.96) belongs to the system with the highest FID; real panoramas sit mid-cohort (15.91). That is the same confound we charge the boundary metrics with, in a different coordinate: it must be read within matched content — the paired protocol of §5.1, where prompts, seeds, resolution and decoder are fixed — and not as a leaderboard column. Such pairing reduces but does not eliminate content confounding: different samplers can generate different scenes from the same prompt and seed. Two of these generators, MVDiffusion and UniPano, additionally send the raw statistic past 10^6 on a few outputs whose profile is flat (MVDiffusion writes blank canvas above and below the panorama strip, so its MAD reaches zero); we floor the MAD at 10^{-3} of the profile median and report medians over images, which leaves every number above unchanged. A learned-feature front end in place of the Scharr response, and a read-out that survives unmatched content, are left for future work.

Used as intended, on the matched pair of §5.1 (identical prompts, seeds, resolution, padding and decoder; one chart against two, neither with PVA), the scan does not resolve a difference on this cohort: RingCut 1.1811 against 1.1767 and RingCut-P 20.26 against 20.13, both with paired 95% bootstrap intervals containing zero over n=1,077. The stored-cut metrics separate the same pair decisively (SeamL1/L2/gradient by $36.7\%/36.0\%/37.6\%$). This is the in-distribution regime of §5.1, where the prior's own seam is weak and a global scan over all longitudes has little to find; it bounds what the yaw-invariant scan adds here, and is why we report it as a guardrail against relocation rather than as the primary seam evidence.

H Diagnostic calibration

Table 13 records controlled-corruption checks for the learned viewport diagnostics. Their response to intended geometry corruptions supports their use as guardrails.

I Extended related work

Text-to-panorama generation. MVDiffusion jointly generates overlapping perspective views with correspondence-aware attention, PanFusion couples perspective and panorama branches through projection-aware interaction, and SMGD introduces spherical-manifold operators (Tang et al., 2023; Zhang et al., 2024; Sun et al., 2025). StitchDiffusion revisits a stitch region during denoising, while UniPano studies LoRA adaptation for text-to-360 generation (Wang et al., 2024a; Ni et al., 2025). DiT360 hybrid-trains a FLUX-style transformer on perspective and polished panoramic data; its released high-resolution model is our frozen base denoiser (Feng et al., 2026). TanDiT uses a sphere-covering tangent-plane grid as its generation representation (Çapuk et al., 2025). In con-

trast, our state remains one full ERP latent; tangent views enter only the training objective for the post-denoiser PVA residual. SpheRoPE modifies spherical rotary position encoding and uses three-way semantic-distortion guidance (Hirschorn et al., 2026). SHERPA combines frequency-selective Circular RoPE and circular latent encoding/decoding with trained image-side FFN adapters and paired/unpaired objectives (Kang et al., 2026). These approaches change token compatibility or internal image-token representations. AtlasFlow leaves the released positional encoding and text condition unchanged and acts on aligned full-ERP velocity predictions. PVA likewise operates outside the Transformer blocks, learning a residual on the denoiser's output rather than adapting hidden image-token features.

Perspective priors and residual adaptation. Latent diffusion models and diffusion transformers inherit strong local image priors from perspective-image training (Rombach et al., 2022; Peebles & Xie, 2023; Black Forest Labs, 2024). Direct ERP inference nevertheless asks a planar token grid to represent nonuniform spherical neighborhoods. PVA uses the same frozen panorama-capable denoiser as the student backbone and training-only teacher: the student predicts on the full ERP, while the teacher receives a matched noisy tangent latent with ordinary square-grid coordinates. The adapter distills their local velocity difference, without a separate teacher model or RGB input at inference. AtlasFlow remains independently applicable when this learned correction is disabled.

Sampling-time consistency and equivariance. MultiDiffusion fuses overlapping diffusion paths on a common canvas (Bar-Tal et al., 2023); SyncDiffusion synchronizes them perceptually, and SpotDiffusion shifts denoising windows (Lee et al., 2023; Frolov et al., 2025). PanoDiffusion and 360DVD rotate latent representations over time, whereas Diffusion360 blends horizontal boundaries after denoising updates (Wu et al., 2024; Wang et al., 2024b; Feng et al., 2023). Our temporal half-turn control represents this rotating-latent family using one model call per step. AtlasFlow instead evaluates alternate full-ERP charts at the same timestep, aligns their complete velocity predictions in one physical frame, and fuses them before one shared update. It neither maintains separate generation trajectories nor overwrites a latent boundary after the update. Evaluation practice varies widely within this training-free family: some methods are scored on a reference panorama collection and report Fréchet distances (Hirschorn et al., 2026), while SphereDiff evaluates twenty scene concepts, enriched with a language model, and LF-MultiDiffusion reuses those prompt sets with multiple seeds (Park et al., 2026; Hayakawa et al., 2026). Their published reference-free scores cannot be inserted directly into a reference-based table; this does not prevent evaluating their methods on a common cohort. We use two reference collections, Matterport3D and ODI-SR1200, and prompt-only open-domain diagnostics.

SphereDiff maps perspective predictions through a spherical latent and applies distortion-aware averaging; LF-MultiDiffusion handles general linear projections through regularized least squares and an iterative solver (Park et al., 2026; Hayakawa et al., 2026). Our longitude-chart maps are lossless integer permutations of the entire ERP, so aligned fusion has a pointwise closed form and uses two full-panorama queries to an already panorama-capable prior. Weighted multi-view aggregation itself is therefore not new; the specific cut-dependent weights and their coordinate behavior are the subject of our equal-compute controls. Generic constructions impose equivariance through symmetrization or transformed inverse mappings (Zhang et al., 2025; Wu et al., 2026a). AtlasFlow specializes this principle to ERP chart cuts, assigning spatial responsibility according to distance from each artificial boundary. The complete finite-chart symmetric construction, our default, yields the guarantee of Sec. 4.1; the local-handoff variant deliberately trades that guarantee for preservation of the primary field away from its cut.

Panorama assessment. Blind omnidirectional IQA assesses perceptual quality without a pristine reference. RL-ScanIQA uses viewport scanpaths. PanoDecouple separates panorama distortion from content and uses Distort-FID for reference-based distribution assessment (Wang et al., 2026; Zheng et al., 2025). Our local geometry probes complement rather than replace such quality scores. Perspective Fields predicts local up-vector and latitude cues and Depth Anything V2 supplies a monocular depth prior (Jin et al., 2023; Yang et al., 2024a). We use these as separate frozen probes of camera plausibility and overlapping-view structure. None of these evaluators supplies the PVA training target.

J PROOF OF PROPOSITION 1

By Eq. (1), shifting the physical latent by $h \in C_K$ rolls every physical-frame prediction by ρ_h and permutes its chart index: $v_g(\rho_h z, t, c) = \rho_h v_{gh}(z, t, c)$. The assumed weight transformation $w_g(\rho_h x) = w_{gh}(x)$ applies the same permutation to the partition. Substituting both into Eq. (4) and re-indexing the finite sum over $G = C_K$ gives $\bar{v}(\rho_h z, t, c) = \rho_h \bar{v}(z, t, c)$. An elementwise Flow-Match Euler update commutes with ρ_h, so induction over the scheduler steps preserves the relation along the complete trajectory. The proof makes no equivariance assumption on F_θ; uniform $1/K$ averaging satisfies the weight condition as well, which is why cut reliability must be tested separately. The reported PVA contains only horizontally circular convolutions, pointwise operations, and roll-invariant normalization, so it commutes with integer cyclic rolls; applying it to an equivariant fused velocity and the correspondingly rolled noisy latent therefore preserves the relation, and the complete method inherits the guarantee. Table 7 verifies the zero packed-latent half-turn error on all eight test cases. Here C_K is represented by integer packed-column rolls, requiring $K \mid W_p$ for uniformly spaced charts. Exactness is an algebraic statement; floating-point kernels may introduce round-off. A stochastic scheduler would additionally require its injected noise to be coupled by the same roll.

K THE LAYOUT AS A SAMPLER COORDINATE: FULL DERIVATION

This section gives the full derivation summarized in Sec. 4.1. The fixed-chart guarantee of Proposition 1 concerns a subgroup of longitude shifts. To distinguish it from changing the coordinate origin of the whole sampler, let R_a roll by $a \in \mathbb{Z}_{W_p}$ packed columns and let $L = (a_1, \ldots, a_K)$ denote physical cut positions. A cut at a_i corresponds to the input roll $-a_i$ in Eq. (1). Writing $\widetilde{F}$ for the frozen denoiser including padding and cropping, define

$$U_a(z, t, c) = R_a \widetilde{F}(R_{-a} z, t, c), \qquad V_L(z, t, c) = \sum_i w_i^L \odot U_{a_i}(z, t, c). \tag{15}$$

For either the symmetric or local profile, moving its weight map with its cut gives $w_i^{L+b} = R_b w_i^L$. Direct substitution therefore yields

$$V_{L+b}(R_b z, t, c) = R_b V_L(z, t, c), \qquad b \in \mathbb{Z}_{W_p}. \tag{16}$$

Indeed, $R_{a_i+b}\widetilde{F}(R_{-a_i-b} R_b z, t, c) = R_b U_{a_i}(z, t, c)$; the weights transform in the same way. No equivariance of the backbone is required. Text conditioning is unchanged; spatial conditioning, when present, must co-rotate. The periodic PVA and an elementwise deterministic scheduler preserve this identity in exact arithmetic. It describes joint input–layout covariance on the latent grid, not arbitrary continuous yaw or exact decoded-RGB equivariance.

Alternatively, draw δ uniformly from $\mathbb{Z}_{W_p}$, independently of roll-invariant initial noise, and keep $L + \delta$ fixed through the sample's trajectory. The joint identity and a change of variables $(z_T, \delta) \mapsto (R_b z_T, \delta + b)$ imply equivariance in distribution. This is a random-rotation wrapper, not a new general symmetrization principle. It adds no denoiser calls to a fixed K-chart sampler. It does not give per-sample equivariance if the input is rolled while the drawn layout stays unchanged. Nor does it by itself remove a defect: with one chart it can simply relocate the seam. Quality gains from fusion remain empirical and must be distinguished from these coordinate identities.

L SEED REPLICATION OF THE MAIN CONFIGURATION

Table 14 repeats the Matterport3D evaluation of the complete method with two further global seeds (every per-image seed is offset by the global seed), for the fixed layout used in the main tables and for the randomized layout of Sec. 4.1. Across the three fixed-layout seeds FID spans 36.20–36.30 and FAED 3.06–3.11, and every seed improves seven of the nine re-evaluated DiT360 metrics. The randomized layout spans 35.89–36.58 in FID. The stored-cut diagnostics move far less than the gap to DiT360 – SeamL1 spans 0.0151–0.0158 across the six runs against 0.0225, DS spans 2.96–3.02 against 3.41, and VPFD 5.53–5.68 against 5.75.

Method	FID↓	FID_{crop} ↓	FID_{pole} ↓	FID_{equ} ↓	FAED↓	IS↑	CS↑	BRISQUE↓	NIQE↓	SeamL1↓	DS↓	VPFD↓
DiT360 (seed 0)	37.90	36.41	62.3	32.30	3.567	5.359	0.3436	22.93	4.116	0.0225	3.412	5.75
Ours, fixed layout, seed 0 (main tables)	36.30	34.43	62.2	31.75	3.114	5.261	0.3433	20.99	4.050	0.0156	2.995	5.63
Ours, fixed layout, seed 1	36.20	34.01	62.9	32.17	3.059	5.275	0.3440	20.59	4.051	0.0158	3.022	5.53
Ours, fixed layout, seed 2	36.27	34.52	63.7	31.66	3.067	5.346	0.3439	20.73	4.062	0.0151	2.964	5.61
Ours, random layout, seed 0	36.58	34.62	62.6	31.54	3.176	5.231	0.3429	21.11	4.054	0.0151	2.993	5.68
Ours, random layout, seed 1	36.58	34.43	63.2	32.16	3.080	5.213	0.3440	20.72	4.055	0.0155	3.011	5.58
Ours, random layout, seed 2	35.89	34.31	63.1	31.55	3.068	5.264	0.3438	20.83	4.063	0.0152	2.974	5.64

Table 14: **Seed replication of the complete method on the Matterport3D cohort.** Within each layout group, prompts, checkpoint, sampler settings, and the eight-column circular decoder are fixed; only the global seed changes. The DiT360 row is one baseline run, not a baseline replication at every seed. Fixed-layout seed 0 is the main configuration. KID is not included in this seed sweep.

Method	n	DS	DS after roll	RingCut	RingCut after roll
Matterport3D cohort prompts, 2048×1024					
DiT360, released inference (planar decoding)	1077	16.73	2.97	1.075	1.075
DiT360, single chart, random cut per sample	1077	2.99	2.95	1.046	1.046
DiT360, symmetric AtlasFlow, fixed layout	1077	2.98	2.95	1.044	1.044
DiT360, symmetric AtlasFlow, random layout	1077	2.98	2.96	1.047	1.047
Matterport3D cohort prompts, 512×1024 (zero-shot baselines at native resolution; DiT360 rows resampled)					
FLUX.1-dev, no panorama handling	1077	128.89	3.33	1.743	1.743
SpheRoPE (zero-shot FLUX.1-dev)	1077	2.67	2.85	1.218	1.218
SphereDiff (zero-shot FLUX.1-dev), 45-prompt subset	45	2.72	2.51	0.787	0.787
DiT360, released inference, resampled	1077	13.74	3.03	1.219	1.219
DiT360, single chart, random cut, resampled	1077	3.10	3.04	1.182	1.182
DiT360, symmetric AtlasFlow, random layout, resampled	1077	3.08	2.99	1.186	1.186
ODI-SR1200 prompts, 2048×1024					
DiT360, per-step padding, circular decoding	1200	3.62	3.43	0.971	0.971
SpheRoPE, official	1200	3.08	3.08	0.975	0.975
DiT360, AtlasFlow local handoff, fixed layout	1200	3.52	3.35	0.961	0.961
Real panoramas, Matterport3D cohort references	1077	1.24	(same)	0.913	(same)
Real panoramas, ODI-SR	1200	7.91	(same)	0.979	(same)

Table 15: **DS at the frame boundary versus after a 180° roll of every image**, and the yaw-invariant RingCut (coherent top-1%) before and after the same roll. A value of one is not a calibrated no-seam threshold. Lower is better for all columns.

M FIXED-LONGITUDE SEAM METRICS CAN BE PASSED BY A ROLL

The discontinuity score DS of Christensen et al. (2024) and the seam-strip metrics used across the panorama literature read the image at one longitude, the frame boundary. Table 15 rolls every generated image by 180°, a lossless relabeling of longitude that changes nothing about the panorama, and re-evaluates. For every method that has a boundary seam, DS collapses to the level of the seamless samplers, including a plain text-to-image model with no panorama handling at all, because the seam is now at the image centre where DS does not look; for methods that are seamless the roll changes nothing, as it should. The yaw-invariant RingCut statistic is unchanged by the roll to all printed digits. The real-panorama rows show a second problem: DS assigns real photographs a worse score than every generator on ODI-SR and a better one on Matterport3D, so its absolute value has no cross-dataset meaning, whereas RingCut places both real sets near one on a common scale. RingCut depends on resolution, so the 512×1024 block compares the zero-shot baselines with the DiT360 arms resampled to the same size. The single-chart random-cut sampler is the per-sample version of the same effect: its DS equals that of the fused samplers because its cut is never at the boundary, although a localized test at its logged cut column still finds the cut on most images. All DiT360 rows in this table use AtlasFlow without PVA and the circular decoder; "released inference" is the public DiT360 pipeline with one-time duplicate-token padding and planar decoding.

N CHART FUSION ON OTHER FROZEN SAMPLERS

Everything in Secs. 4.1 and 4.1 is a statement about coordinates, not about a network, so the same code wraps any sampler that predicts a velocity on an ERP grid: roll the latent and any spatial condition into each chart, evaluate the frozen model, roll back, fuse, and move the cuts with the input.

Frozen sampler	Quality (base → wrapped)	One-step flow error	Placement drift
DiT360	FID 59.64 → 55.18; FAED 16.82 → 12.79 (ODI-SR1200)	0.09–0.14 → 0	RGB 0.14 → 0.009; latent 0.42 → 0
SpheRoPE	FID 74.70 → 73.67; OmniFID 95.64 → 94.86 (ODI-SR1200)	Not measured	RGB 0.22 → 0.006; latent 0.46 → 0
HunyuanWorld-1.0	$\text{PSNR}_{\text{free}}$ 9.48 → 9.53; LPIPS 0.716 → 0.716 (ground truth)	Not measured	RGB 0.31 → 0.22; 46/48 pairs improve

Table 16: **The same fusion rule on three frozen samplers**, without PVA. Quality: the metric the base method reports, before and after wrapping, one protocol per row. One-step error: flow equivariance under a yaw shift of the noise (one denoising step, bitwise comparison); zero means bitwise equality in that test. Placement drift: relative RMSE of the completed region when the same real photograph is pasted at different longitudes and the results are unrolled; zero means the completion does not depend on where the photograph was pasted in that representation. SpheRoPE's latent placement error is not a one-step flow measurement. HunyuanWorld additionally changes its final blending policy; the no-blend control is reported below.

Table 16 applies the symmetric two-chart rule with a randomized layout, and no PVA, to two further frozen priors of different kinds, a zero-shot positional-encoding method on the untuned FLUX.1-dev (SpheRoPE) and an image-conditioned outpainter (HunyuanWorld-1.0, FLUX.1-Fill with a LoRA), and reports what the wrapper does to the metric the base method reports, whether the co-rotated sampler is exactly equivariant, and how much a completion changes when the conditioning photograph is pasted at a different longitude (*placement drift*: relative RMSE of the completed region after unrolling, over real photographs pasted at several headings). DiT360 uses 16 photographs at eight headings (112 nonzero-heading pairs), SpheRoPE uses eight at four headings (24 pairs), and HunyuanWorld uses 16 at $0, 45, 90, 180$ degrees (48 pairs). For the first two, the placement comparison holds two-chart fusion fixed and changes only whether the layout co-rotates; it is not a one-chart versus two-chart test.

Three readings follow. First, the trained prior improves the listed distribution metrics, while the outpainter changes little. Second, the coordinate identity is independent of the prior, but its numerical tests must be distinguished. A separate eight-sample SpheRoPE noise-roll probe has decoded-RGB relative error $0.4528 \to 0.0059$ after complete sampling; the $0.46 \to 0$ in the table instead measures conditional final-latent drift. Third, pasting the same photograph at a different longitude produces mean completed-region RGB relative RMSE of 0.14–0.31 in the reference arms; this is not a percentage of pixels or scene content changed. For the two samplers whose full pipeline we control the wrapped drift is zero in latent space and under 1% in pixels (the remainder is the decoder); for HunyuanWorld, whose image-conditioning pipeline does not obey the full transformation law, drift falls by 28% (46 of 48 photograph–heading pairs improve). This comparison also disables the released final blend. Against a no-blend reference, drift falls from 0.272 to 0.221 and 39 of 48 pairs improve, so final blending alone does not explain the reduction. The experiment does not isolate the source of all residual drift. Fidelity to the ground truth is a separate question on which the trained outpainter remains best (about 0.5 dB, 0.03 SSIM, and 0.02 LPIPS better than completions driven by the text-to-panorama prior with a clamped view); we claim the placement property, not completion quality. The completion protocol cuts an $80°$ perspective view from a real ODI-SR panorama, pastes it at longitude θ into an empty ERP canvas, completes the rest, unrolls by $-\theta$, and compares with the ground truth outside the pasted region (PSNR on the free pixels, LPIPS at 1024×512).

O ADDITIONAL QUALITATIVE RESULTS

All panoramas in this section are rolled so that the stored cut is at the centre of the ERP; perspective views use a $90°$ square camera and their footprints are drawn in **blue**. Figure 8 shows ten of the twelve systems of Table 1 (all but Diffusion360 and PanFusion) on three cohort prompts.

Gallery. Figure 7 shows twelve outputs of the complete method (AtlasFlow with PVA on frozen DiT360): six ODI-SR1200 prompts and six prompts from OOD1000, a set of 1,000 stratified out-of-domain prompts (night streets, rooftops, forests, industrial interiors, invented scenes) with no

reference panoramas, chosen by inspection for scene variety among the 48 top-ranked images of each cohort under a rank average of per-image CLIPScore, SeamL1 and VPFD.

How the figure examples were chosen. The two panels of Figure 4 were selected with the per-image metric that the corresponding component targets; the final pick among the top-ranked candidates was made by inspection. For panel (a) we ranked all cohort panoramas by the SeamL1 reduction from PVA-only to the complete method and kept pairs whose two arms generate the same scene (RGB 128×64 thumbnail correlation above 0.90); the panel (a) example is rank 73 of the roughly 1k cohort panoramas. For panel (b) we ranked all 17,232 evaluated viewports by the VPFD reduction from AtlasFlow-only to the complete method, discarded viewports whose baseline VPFD exceeds 35° (on nearly featureless views such as blank walls the Perspective Fields probe returns arbitrary values, so a large "improvement" there is a probe artifact rather than a geometry fix) and viewports without measurable structure in either arm; the panel (b) view (yaw -180°, pitch -45°) is the second-ranked survivor. Because PVA changes the velocity at every step, the two arms do not produce pixel-identical scenes even at a fixed seed; fine details therefore differ within the same room.

AI USE STATEMENT

Generative AI tools assisted implementation and debugging, result preparation and interpretation, literature search, translation, and the drafting and polishing of the manuscript. All reported numbers come from model outputs and explicit evaluation procedures.

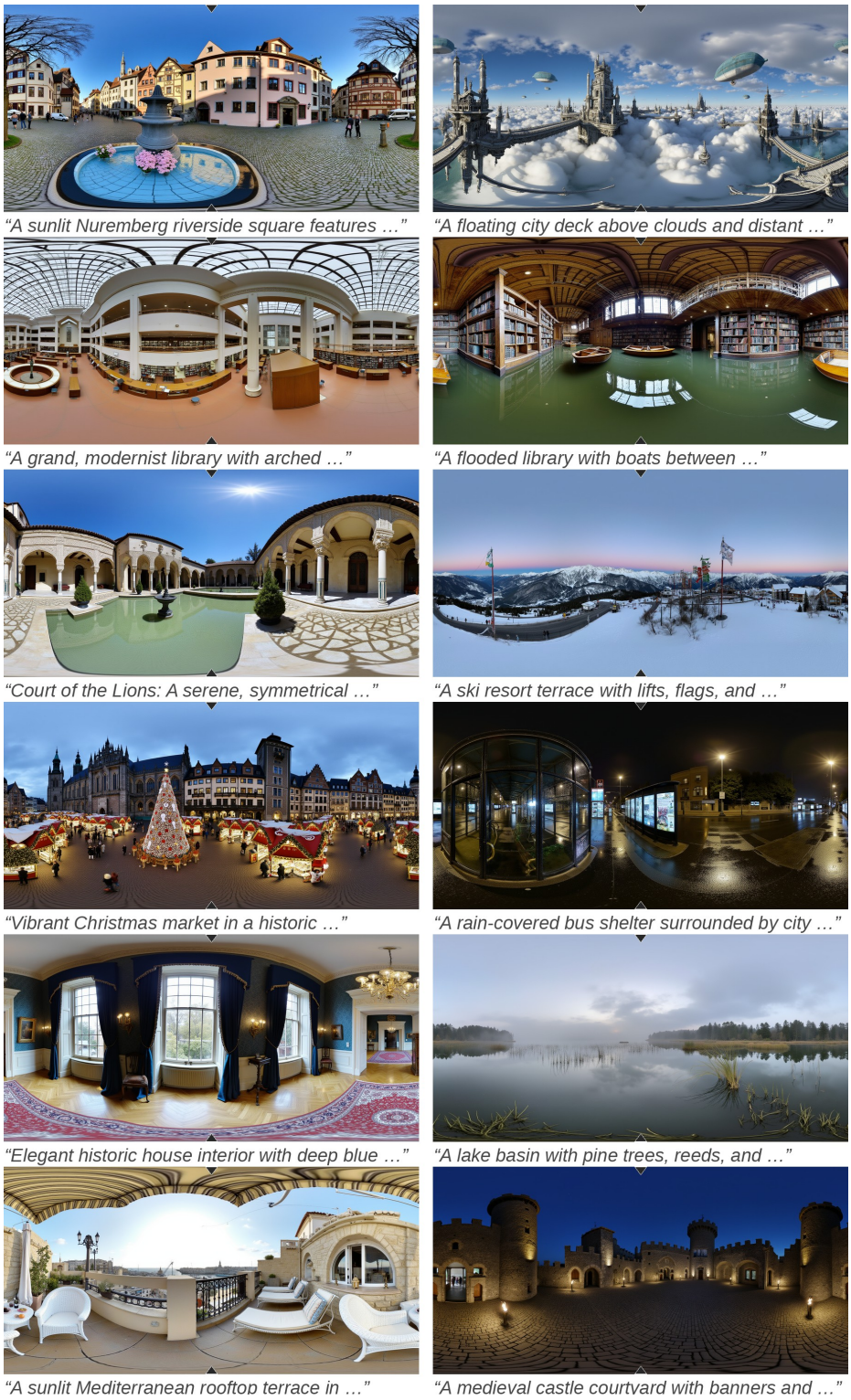

Figure 7: **Gallery of the complete method on twelve prompts**: left column ODI-SR1200, right column OOD1000, abbreviated prompt under each panorama. Every ERP is rolled so that the stored cut is the centre column (marked at the top and bottom edges). Chosen by inspection for variety among the 48 top-ranked images of each cohort (per-image CLIPScore, SeamL1, and VPFD). Appearance is the frozen backbone's; what AtlasFlow and PVA change, the closed cut and the tangent-view geometry, is measured in Table 10 and shown at camera scale in the figures that follow.



Figure 8: **Ten of the twelve systems of Table 1 (all but Diffusion360 and PanFusion) on three cohort prompts**, each at its released resolution, stated under its name. MVDiffusion's released pipeline pads the latitudes outside its stitched views. Ours is the main-table configuration.