PATCH-SHEAF RESIDUAL MATCHING FOR LOCAL-TO-GLOBAL COHERENCE IN LATENT DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern image generators produce locally plausible content while distorting structures that extend across spatial regions, such as roads, lanes, and object boundaries. We formulate this failure as a local-to-global compatibility problem over predicted-clean latent patch states. We construct a *cellular sheaf* with neighbouring-patch, parent-region, and sliding square-cell relations, whose residuals quantify incompatibility between local states. Clean latents can exhibit meaningful nonzero relation residuals at boundaries and transitions, motivating the use of clean relation energies as matching targets. We therefore match predicted-clean relation energies, *i.e.*, the squared norms of the residuals, to clean-latent targets, calibrate per-patch incident-energy distributions, and weakly preserve their spatial organization. These auxiliary losses operate only within an intermediate signal-to-noise ratio window; the +tail extension adds timestep-conditioned tail calibration and localization, while the generative objective remains active at all timesteps. Using a DiT-XL/2 backbone, we evaluate on DeepGlobe Roads, LSUN-Church, and BDD100K against no-sheaf controls matched for training duration and learning-rate schedule. Across five independent phase-2 training seeds, +tail reduces mean FID relative to staged-LR controls by 11.8%, 12.9%, and 21.4%, respectively, while consistently improving KID, CMMD, recall, and coverage.

1 INTRODUCTION

Diffusion-based generative models achieve high-fidelity image synthesis across unconditional and latent-generation settings (Dhariwal & Nichol, 2021; Rombach et al., 2022). However, perceptual quality does not necessarily imply coherent spatial structure. Generators can produce locally plausible regions while distorting structures that extend across them, including roads, lane markings, facades, perspective-aligned contours, and object boundaries. Existing benchmarks likewise expose weaknesses in spatial relations, attribute binding, and complex scene composition (Huang et al., 2023). Window-based generation further demonstrates that locally seamless regions can form globally inconsistent outputs when coordination is insufficient (Lee et al., 2023). These observations identify a general local-to-global problem: plausible local content does not guarantee coherent global assembly. Here, *local-to-global* refers to the structural consistency propagated across multiple adjacent latent patches through overlapping local relations; it does not imply explicit modelling of arbitrary scene-wide semantic dependencies. Prior work addresses different aspects of this problem through global-context transformer backbones (Bao et al., 2023; Peebles & Xie, 2023), cascaded or hierarchical generation (Ho et al., 2022), and synchronization across spatial windows (Bar-Tal et al., 2023; Lee et al., 2023). Other methods introduce explicit spatial or topological supervision (Zhang et al., 2023; Gupta et al., 2025). For example, Glob-Diffusion uses hierarchical guidance and a region-guided adapter to improve large-scale composition (Kang et al., 2026), whereas TopoDiffusionNet uses persistent-homology objectives to encourage preservation of connected components and holes (Gupta et al., 2025). These approaches demonstrate the value of structural guidance, but they often lack the cross-patch structural coherence, which can be effectively extracted by explicitly representing a predicted-clean latent as a collection of local states. We investigate this compatibility using cellular sheaves. A cellular sheaf assigns local state spaces to cells and restriction maps to their relations, thereby specifying when local assignments agree (Curry, 2014; Hansen & Ghrist, 2019). A globally compatible assignment forms a section, while the cellular coboundary records violations of the local compatibility conditions. The associated sheaf-Laplacian quadratic

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.

form summarizes their aggregate magnitude (Hansen & Ghrist, 2019). Cellular sheaves have also been used to represent feature compatibility and diffusion in neural systems (Bodnar et al., 2022). This formalism is therefore well suited to image generation, where latent patches encode local information but must assemble into a coherent object, scene, or spatial layout. We formulate latent generation as a patch-gluing problem. Given a predicted-clean latent, we organize its spatial patches into neighboring-edge, parent-region, and local square-cell relations. Their sheaf residuals quantify local incompatibility, which can reflect meaningful transitions as well as structural distortions such as interrupted roads or misaligned architectural elements (Figure 7). Prior work relates oversmoothing in graph neural networks to sheaf geometry (Bodnar et al., 2022), while diffusion-based generation and editing highlight the importance of edge preservation (Vandersanden et al., 2025; Gong et al., 2026). Under our patch relations, clean latents can retain meaningful nonzero residuals at boundaries and transitions. We therefore match predicted-clean relation energies to corresponding clean-target values, whether zero or nonzero. We test zero-residual minimization as a controlled alternative in Appendix 5.6, where clean-target matching provides better residual calibration. Complementary losses calibrate the distribution of incident patch energies and weakly preserve their spatial organization, while a +tail extension targets high-residual calibration and localization across denoising regimes. These losses operate as signal-to-noise ratio (SNR)-gated auxiliary objectives and do not replace the standard diffusion objective.

Our contributions are: **(i)** a cellular-sheaf representation of predicted-clean latent patches for modelling cross-patch compatibility, using adaptive neighbouring-edge restrictions and fixed parent-region and square-cell relations; **(ii)** a data-calibrated objective that matches relation energies to clean targets while preserving meaningful nonzero residual structure, complemented by patch-distribution and spatial matching, with a +tail extension for timestep-conditioned calibration and localization; and **(iii)** improvements in generation quality and structural coherence over no-sheaf controls, showed by experiments on the DeepGlobe Roads, LSUN-Church, and BDD100K datasets.

2 METHOD

2.1 OVERVIEW OF TRAINING PROTOCOL

Training proceeds in two stages. In *phase 1*, we train a DiT-XL/2 latent diffusion model (Peebles & Xie, 2023) from scratch using only the standard generative objective $\mathcal{L}_{\text{gen}}$. For each dataset, the checkpoint with the lowest Fréchet Inception Distance (FID) initializes all phase-2 patch-sheaf branches and two no-sheaf controls: a duration-matched continuation and a continuation additionally matched to the staged learning-rate schedule. All remaining image-quality metrics are computed after selection under the fixed evaluation protocol; Section 2.5 gives the complete comparison protocol. In *phase 2*, $\mathcal{L}_{\text{gen}}$ remains active at every timestep. Given a clean latent x_0, its noisy realization z_t, and timestep t, the model predicts noise $\hat{\epsilon}_\theta(z_t, t)$; standard diffusion inversion yields the predicted-clean latent $\hat{x}_{0,\theta}(z_t, t)$, abbreviated below as $\hat{x}_0$. Within an intermediate signal-to-noise ratio (SNR) window, three auxiliary blocks compare $\hat{x}_0$ with x_0: relation-level clean-target energy matching, patch-level residual-distribution calibration, and spatial patch-residual matching. Figures 1 and 2 summarize the overall principle and residual representations, respectively.

2.2 PATCH-SHEAF CONSTRUCTION

Patch complex and sheaf residuals. We construct a finite patch complex over non-overlapping latent patches obtained by average pooling. Let V denote the patches, E the neighbouring patch pairs, P the non-overlapping parent regions formed by grouping patches, and C the sliding 2×2 square cells. Horizontal, vertical, and diagonal edges form an eight-neighbour patch graph. Each patch $v \in V$ carries a state x_v, where $x \in \{\hat{x}_0, x_0\}$. For an edge $e = (u, v)$, restriction maps $\rho_{u\to e}$ and $\rho_{v\to e}$ send x_u and x_v to a shared edge space, defining

$$r_e(x) = \rho_{u\to e}(x_u) - \rho_{v\to e}(x_v). \tag{1}$$

Parent and cell residuals analogously compare each participating patch state with its mean parent or cell state. The channel-averaged edge-residual energy is $\varepsilon_e(x) = \|r_e(x)\|_2^2/d$, where d is the latent-channel dimension.

Residual operator. Once the restriction maps are fixed or the adaptive maps are instantiated for a sample, let R_A, $A \in \{E, P, C\}$, denote the resulting linear residual operator, such that $R_A(x)$

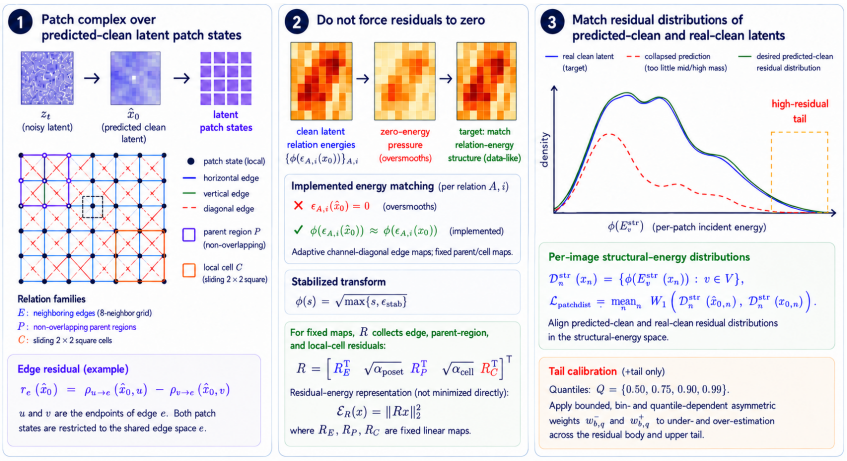

Figure 1: Patch-sheaf residual matching. Relation-level, distributional, and spatial residual structures are matched to clean targets; +tail adds timestep-conditioned tail calibration and localization.

stacks the relation residuals $r_{A,i}(x)$. We combine these operators as

$$R = \left[R_E^\top \quad \sqrt{\alpha_{\text{poset}}}\, R_P^\top \quad \sqrt{\alpha_{\text{cell}}}\, R_C^\top\right]^\top. \tag{2}$$

The coefficients α_{poset} and α_{cell} scale the parent and cell components relative to the unit-weight edge component. The three blocks share these relation families but use different restriction maps and aggregations (Figure 2). For fixed pairwise restrictions, $R_E^* R_E$ is the corresponding sheaf Laplacian (Hansen & Ghrist, 2019; Bodnar et al., 2022).

Adaptive restriction maps. Block I uses feature-conditioned diagonal edge maps $\rho_{v\to e}(x) = \text{diag}(\mathbf{1} + \alpha_\rho \tanh(g_{\theta,v\to e}(x)))$, where $\alpha_\rho > 0$ bounds the channel-wise deviation from the identity. Parent and cell maps remain fixed averages. The adaptive maps are regularized toward identity and induce the conditional Laplacian $\Delta_\theta(x) = R_{\theta,x}^* R_{\theta,x}$; further details are provided in Appendix 5.1.

2.3 Three-Block Auxiliary Patch-Sheaf Objective

Figure 2 summarizes the three auxiliary blocks. They share the $E/P/C$ relations but use different representations: Block I matches individual relation energies using adaptive edge maps and fixed parent/cell maps; Block II aggregates separately computed fixed-map residuals by patch; and Block III augments these patch energies with spatial coordinates. Below, we describe these blocks in detail. For convenience, n indexes mini-batch samples, $(z)_+ = \max(z, 0)$, and the λ's denote nonnegative loss weights throughout the three blocks.

Block I: relation-level geometric compatibility. For $A \in \{E, P, C\}$, let $\varepsilon_{A,i}(x)$ denote the channel-averaged energy of relation instance i. For edge and parent relations, $\varepsilon_{A,i}(x) = \|r_{A,i}(x)\|_2^2/d$; for a square cell, it is the mean of its four patch-to-cell energies. Using the stabilized transform $\phi(s) = \sqrt{\max\{s, \epsilon_{\text{stab}}\}}$, where $\epsilon_{\text{stab}} > 0$ is fixed, we match corresponding predicted-clean and clean energies:

$$\mathcal{L}_A^{\text{match}} = \text{mean}_i \left[\phi(\varepsilon_{A,i}(\hat{x}_0)) - \phi(\varepsilon_{A,i}(x_0))\right]^2. \tag{3}$$

By treating the clean energies as stop-gradient targets, the complete objective is

$$\mathcal{L}_{\text{geom}} = \lambda_{\text{sheaf}}\left(\mathcal{L}_E^{\text{match}} + \alpha_{\text{poset}}\mathcal{L}_P^{\text{match}} + \alpha_{\text{cell}}\mathcal{L}_C^{\text{match}} + \lambda_\rho \mathcal{L}_\rho\right) + \lambda_{\text{stat}}\mathcal{L}_{\text{stat}}. \tag{4}$$

Here, $\mathcal{L}_\rho$ penalizes the outputs parameterizing the adaptive edge restrictions, keeping the maps near identity and discouraging them from trivially cancelling relation residuals. The term $\mathcal{L}_{\text{stat}}$ matches aggregate patch-state statistics; complete definitions are given in Appendix 5.1.

Block II: patch-level residual-energy calibration. Block II independently recomputes residuals using identity edge maps and the fixed parent and cell constructions; it does not aggregate Block I's

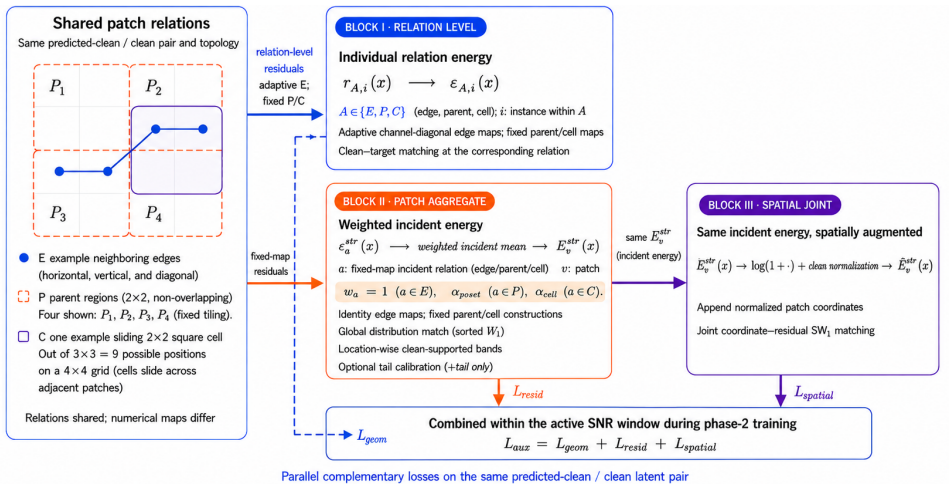

Figure 2: Block-specific residual representations: relation-level matching (I), fixed-map patch aggregation (II), and spatial augmentation (III).

adaptive edge energies. For each fixed-map relation a, let $\varepsilon_a^{\mathrm{str}}(x)$ denote its energy, using the same channel normalization and cell-averaging convention as Block I. For the relations $\mathcal{I}(v)$ incident to patch v, define

$$E_v^{\mathrm{str}}(x) = \frac{\sum_{a\in\mathcal{I}(v)} w_a \varepsilon_a^{\mathrm{str}}(x)}{\sum_{a\in\mathcal{I}(v)} w_a}, \tag{5}$$

where $w_a = [1, \alpha_{\mathrm{poset}}, \text{or } \alpha_{\mathrm{cell}}]$ for edge, parent, or cell relations, respectively.

We first match the transformed incident-energy distributions independently for each image. Defining $\mathcal{D}_n^{\mathrm{str}}(x) = \{\phi(E_v^{\mathrm{str}}(x_n)) : v \in V\}$, we use

$$\mathcal{L}_{\mathrm{patchdist}} = \mathrm{mean}_n\, W_1\big(\mathcal{D}_n^{\mathrm{str}}(\hat{x}_0), \mathcal{D}_n^{\mathrm{str}}(x_0)\big)\,. \tag{6}$$

For these equally weighted one-dimensional distributions, W_1 is computed by sorting the two patch-energy vectors and averaging their absolute differences (Peyré & Cuturi, 2019; Ramdas et al., 2017). Because sorting discards patch location, we additionally use a location-wise band loss. At each patch location, it penalizes predicted energies outside the clean 0.10–0.90 quantile band, with an additional penalty above the clean 0.99 quantile. This discourages local residual collapse and extreme amplification without enforcing exact patchwise correspondence. Its definition is given in Appendix 5.1. The +tail extension also acts on these incident energies (see details in Appendix 5.2). It adds timestep-binned quantile and upper-tail calibration $\mathcal{L}_{\mathrm{calib}}$, mid- and high-residual mode preservation $\mathcal{L}_{\mathrm{mode}}$, and high-residual spatial localization $\mathcal{L}_{\mathrm{loc}}$. We define

$$\mathcal{L}_{\mathrm{resid}} = \lambda_{\mathrm{patchdist}}\mathcal{L}_{\mathrm{patchdist}} + \lambda_{\mathrm{band}}\mathcal{L}_{\mathrm{band}}, \qquad \mathcal{L}_{\mathrm{tail}} = \lambda_{\mathrm{tail}}\mathcal{L}_{\mathrm{calib}} + \lambda_{\mathrm{mode}}\mathcal{L}_{\mathrm{mode}} + \lambda_{\mathrm{loc}}\mathcal{L}_{\mathrm{loc}}. \tag{7}$$

Block III: spatial patch-residual matching. Block III reuses the fixed-map incident energy E_v^{str} from Block II and jointly represents residual magnitude and patch position. For sample n, define $h_{n,v}(x_n) = \log(1 + E_v^{\mathrm{str}}(x_n))$, and let μ_n^{clean} and $\sigma_n^{\mathrm{clean}}$ be the mean and standard deviation of $\{h_{n,v}(x_{0,n}) : v \in V\}$. Using these clean statistics to normalize both members of the pair, we form

$$\widetilde{E}_{n,v}^{\mathrm{str}}(x_n) = \frac{h_{n,v}(x_n) - \mu_n^{\mathrm{clean}}}{\max\{\sigma_n^{\mathrm{clean}}, \epsilon_{\mathrm{stab}}\}}, \qquad p_v(x_n) = \Big[\omega_{xy} s_v^x,\ \omega_{xy} s_v^y,\ \omega_r \widetilde{E}_{n,v}^{\mathrm{str}}(x_n)\Big]\,. \tag{8}$$

where $x_n \in \{\hat{x}_{0,n}, x_{0,n}\}$ and $s_v = (s_v^x, s_v^y) \in [0,1]^2$ is the normalized patch coordinate. The point-space weights ω_{xy} and ω_r balance spatial position and residual magnitude; both are set to one in the reported models. The spatial loss is defined as:

$$\mathcal{L}_{\mathrm{spatial}} = \lambda_{\mathrm{jointpatch}}\, \mathrm{mean}_n\, \mathrm{SW}_1\left(\{p_v(\hat{x}_{0,n}) : v \in V\}, \{p_v(x_{0,n}) : v \in V\}\right), \tag{9}$$

where SW_1 is the sliced Wasserstein-1 distance. Shared random projections are used for each predicted-clean/clean pair, providing weak spatial alignment without enforcing exact patchwise equality (Deshpande et al., 2018; Kolouri et al., 2019).

2.4 SNR-Gated Phase-2 Objective

The auxiliary losses are activated only within an intermediate SNR window. Defining the core and +tail objectives as $\mathcal{L}_{\text{aux}}^{\text{core}} = \mathcal{L}_{\text{geom}} + \mathcal{L}_{\text{resid}} + \mathcal{L}_{\text{spatial}}$ and $\mathcal{L}_{\text{aux}}^{+\text{tail}} = \mathcal{L}_{\text{aux}}^{\text{core}} + \mathcal{L}_{\text{tail}}$, we write

$$w_{\text{sheaf}}(t) = \mathbf{1}\{0.5 \le \text{SNR}(t) \le 10\}, \quad \mathcal{L}^{(m)}(t) = \mathcal{L}_{\text{gen}} + w_{\text{sheaf}}(t)\mathcal{L}_{\text{aux}}^{(m)}, \quad m \in \{\text{core}, +\text{tail}\}. \tag{10}$$

Thus, $\mathcal{L}_{\text{gen}}$ remains active at every timestep, whereas the predicted-clean auxiliary losses operate only within the selected SNR regime. The +tail calibration uses timestep bins within the same window (Nichol & Dhariwal, 2021; Choi et al., 2022; Karras et al., 2022; Hang et al., 2023).

2.5 Model Variants, Controls, and Ablation Protocol

Primary comparisons. All phase-2 branches start from the same selected phase-1 checkpoint, which is retained as the pre-continuation baseline. To assess our main contributions, we compare: *P1 continuation*, trained only with $\mathcal{L}_{\text{gen}}$; *P1 continuation + staged LR*, which uses the phase-2 learning-rate schedule without auxiliary losses; *full patch-sheaf*, which adds $\mathcal{L}_{\text{geom}}$, $\mathcal{L}_{\text{resid}}$, and $\mathcal{L}_{\text{spatial}}$; and *full patch-sheaf + tail*, which additionally includes $\mathcal{L}_{\text{tail}}$. The no-sheaf branches control for continued training and learning-rate scheduling, while the two patch-sheaf variants isolate the tail extension. Principal phase-2 checkpoints are selected using the joint image-quality and residual-calibration profile; matched comparisons evaluate all branches at the same dataset-specific total training step. The matched staged-LR, full-sheaf, and +tail comparison is repeated over five independent phase-2 training seeds; we report the mean and sample standard deviation across seeds.

Objective-component and robustness analyses. Appendix 5.5 reports matched independent branches for the staged-LR no-sheaf control, Block I, Blocks I–II, and Blocks I–III. Blocks are enabled cumulatively across separate runs, not sequentially within one run. Appendix 5.6 further compares clean-target matching with zero-residual minimization at Blocks I, I–II, and I–III. We also evaluate multi-hop structural continuation (Section 3.3), FFHQ domain alignment (Appendix 5.4), BDD100K sensitivity to the LR-shock interval, SNR window, and block weights (Appendix 5.9), and dataset-dependent patch resolution (Appendix 5.10).

2.6 Datasets and Quantitative Evaluation

We evaluate three spatially structured image domains. DeepGlobe Roads (Demir et al., 2018) contains 256×256 RGB satellite images testing road and land-boundary connectivity. LSUN-Church Outdoor (Yu et al., 2015) tests cross-region architectural alignment, while BDD100K (Yu et al., 2020) tests lane, boundary, and perspective geometry in heterogeneous street scenes. Our quantitative evaluation covers *(i)* standard generation quality, *(ii)* latent residual calibration, *(iii)* pixel-space structure, and *(iv)* residual–edge alignment and cross-patch geometric coherence.

2.6.1 Image Quality and Continuation Controls

Under a fixed 50k-sample protocol, we report FID, KID, sFID, CMMD, precision, recall, and coverage. sFID adds spatial sensitivity, CMMD provides a CLIP-based distributional comparison, and coverage measures representation of the real-data manifold (Nash et al., 2021; Jayasumana et al., 2024; Naeem et al., 2020). Phase-2 checkpoints are selected using FID and residual-calibration diagnostics; all other metrics are computed only after selection. Matched comparisons use the same dataset-specific step, with staged-LR, full patch-sheaf, and +tail results averaged over five seeds.

2.6.2 Latent Patch-Residual Geometry

We evaluate residual geometry using fixed clean images, timesteps, and noise samples, enabling paired comparisons across methods and checkpoints. For probe timestep or SNR bin b, let

$$\widehat{\mathcal{D}}_b = \{\phi(E_v^{\text{str}}(\hat{x}_{0,n})) : n \in b,\ v \in V\}, \qquad \mathcal{D}_b^0 = \{\phi(E_v^{\text{str}}(x_{0,n})) : n \in b,\ v \in V\}.$$

ResW1 $\downarrow$	**Mass ratio** $\to 1$	**QCE** $\downarrow$	**TailIoU**$_q$ $\uparrow$
$W_1(\widehat{\mathcal{D}}_b, \mathcal{D}_b^0)$	$\frac{\text{mean}(\widehat{\mathcal{D}}_b)}{\text{mean}(\mathcal{D}_b^0)}$	$\text{mean}_q \lvert\widehat{Q}_{b,q} - Q_{b,q}^0\rvert$	$\mathbb{E}_{n\in b}\,\text{IoU}(\widehat{H}_{n,q}, H_{n,q}^0)$

Here, $\widehat{Q}_{b,q}$ and $Q_{b,q}^0$ are the corresponding residual quantiles, while $\widehat{H}_{n,q} = H_q(\hat{x}_{0,n})$ and $H_{n,q}^0 = H_q(x_{0,n})$ are their high-residual patch sets. We use $\text{IoU}(A, B) = |A \cap B|/(|A \cup B| + \epsilon)$. These diagnostics measure residual-distribution agreement, mass preservation, quantile calibration, and high-residual localization, respectively.

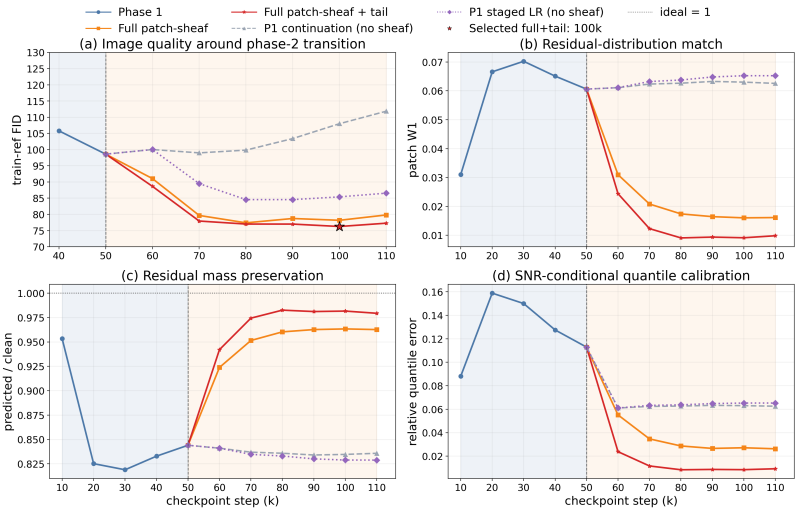

Figure 3: DeepGlobe phase transition. The dashed line marks phase-2 initialization.

Boundary-conditioned residual calibration. To test calibration near visible structural transitions, we compute a Sobel edge map for each clean probe image, pool it to the patch grid, and split patches into boundary and non-boundary regions using edge-strength threshold q_e ($q_e = 0.95$ in the reported results). For each region $r \in \{\mathrm{bdry}, \mathrm{nonbdry}\}$, we restrict $\widehat{\mathcal{D}}_b$ and $\mathcal{D}_b^0$ accordingly and recompute ResW1, QCE, and the residual-mass ratio. This tests whether calibration improves near meaningful image structure rather than only globally.

2.6.3 Pixel-Space Structure

We compare generated and real images using Wasserstein distances over edge density, boundary contrast, 95th-percentile gradient magnitude, colourfulness, and saturation. These diagnostics test whether latent improvements yield more data-like visible structure.

2.6.4 Cross-Space Micro-Level Geometry

We test whether high latent-residual activity aligns with visible edges and whether edge geometry continues coherently across patch boundaries. *Residual-edge alignment* compares the upsampled high-residual mask $H_q(x)$ with the image edge map $\mathcal{E}(x)$. We report precision, recall, and $F_{1,q} = 2P_qR_q/(P_q + R_q + \varepsilon)$, where higher $F_{1,q}$ indicates stronger alignment. For *patch-boundary coherence*, we match high-gradient pixels $i = (a, b)$ on opposite sides of each patch boundary. Let $\theta = \arctan 2(I_y, I_x)$, $\delta_i = \theta_a - \theta_b$, and $w_i = \min(\|\nabla I_a\|, \|\nabla I_b\|)$. We compute

$$\Delta\theta_i = \frac{1}{2}\,|\arctan 2(\sin 2\delta_i, \cos 2\delta_i)| \qquad \text{BAE } (\downarrow) \quad \frac{\sum_i w_i \Delta\theta_i^\circ}{\sum_i w_i} \qquad \text{IPGC } (\uparrow) \quad \frac{\sum_i w_i \cos(2\Delta\theta_i)}{\sum_i w_i} \tag{11}$$

where $\Delta\theta_i^\circ = (180/\pi)\Delta\theta_i$. The axial discrepancy treats opposite gradient directions as equivalent. Our principal metrics are BAE and IPGC: lower BAE and higher IPGC indicate more coherent edge continuation across neighbouring patches.

2.7 Qualitative Evaluation

We provide two supporting qualitative analyses. Figure 5 compares fixed-seed samples from the phase-1 baseline, staged-LR control, full patch-sheaf, and +tail models at matched checkpoints, showing that the proposed patch-sheaf significantly improves the generated geometry consistency. It complements the image-space diagnostics in Table 4. Shared seeds aid comparison but do not imply pixelwise correspondence between independently fine-tuned models. Figure 7 instead uses paired probes with identical clean images, timesteps, and noise to visualize structure across latent-patch boundaries, complementing the BAE and IPGC results in Figure 6. These examples provide qualitative support rather than supervised localization or reconstruction evidence.

3 Results

Following the evaluation hierarchy, we report *(i)* generation quality and matched continuation controls, *(ii)* global and boundary-conditioned latent residual calibration, *(iii)* aggregate pixel-space structure, and *(iv)* residual-edge alignment and cross-patch geometric coherence.

Table 1: Selected checkpoints using 50k generated samples. P1: phase-1 baseline; +T: sheaf + tail. KID is multiplied by 10^3. Bold marks the better phase-2 result.

			Distribution quality			Data support		Additional	
Data	Model	Step	FID ↓	KID ↓	CMMD ↓	Rec. ↑	Cov. ↑	sFID ↓	Prec. ↑
DeepGlobe	P1	50k	97.486	85.90	1.7132	.124	.436	45.634	.4865
	Full	80k	76.245	59.11	**1.6551**	**.189**	**.464**	**41.431**	**.3553**
	+T	100k	**76.093**	**57.73**	1.7662	**.189**	.445	42.427	.3249
LSUN-Church	P1	290k	18.435	15.03	2.5573	.347	.635	31.693	.5376
	Full	390k	10.337	7.13	2.1750	**.446**	.746	25.253	**.5493**
	+T	390k	**10.144**	**6.94**	**2.1605**	.440	**.755**	**25.247**	.5466
BDD100K	P1	110k	12.037	8.83	1.3132	.440	.819	17.644	.5925
	Full	170k	8.614	5.54	1.1926	.470	.880	16.427	.6232
	+T	170k	**8.273**	**5.24**	**1.1909**	**.477**	**.888**	**16.374**	**.6234**

Table 2: Matched results over five seeds (SD in parentheses; 50k samples/seed). P1C: phase-1 continuation; SLR: staged LR; +T: sheaf +tail. KID is multiplied by 10^3. Bold: best per dataset.

		Distribution quality			Data support		Additional	
Data	Model	FID ↓	KID ↓	CMMD ↓	Rec. ↑	Cov. ↑	sFID ↓	Prec. ↑
DeepGlobe	P1C	106.91 (.27)	93.92 (.57)	1.999 (.004)	.107 (.006)	.356 (.004)	49.40 (.42)	**.386 (.003)**
	SLR	86.92 (4.13)	71.60 (5.12)	1.874 (.076)	.167 (.002)	.430 (.008)	**40.75 (2.04)**	.360 (.031)
	Full	77.72 (.19)	59.80 (.08)	1.786 (.002)	.184 (.007)	.444 (.004)	42.47 (.40)	.325 (.006)
	+T	**76.68 (.58)**	**58.50 (.74)**	**1.772 (.005)**	**.186 (.003)**	**.445 (.009)**	42.65 (.28)	.321 (.008)
LSUN-Church	P1C	17.39 (.10)	14.28 (.09)	2.496 (.003)	.368 (.001)	.661 (.003)	30.94 (.13)	.545 (.006)
	SLR	11.79 (.26)	8.52 (.29)	2.276 (.020)	.407 (.001)	.746 (.005)	27.41 (.22)	**.576 (.002)**
	Full	10.40 (.15)	7.27 (.17)	2.182 (.008)	**.442 (.005)**	.749 (.006)	**25.21 (.10)**	.543 (.006)
	+T	**10.28 (.11)**	**7.09 (.15)**	**2.171 (.009)**	.440 (.001)	**.750 (.005)**	25.27 (.12)	.542 (.004)
BDD100K	P1C	11.81 (.05)	8.70 (.06)	1.274 (.004)	.442 (.003)	.826 (.005)	17.58 (.03)	.599 (.003)
	SLR	10.63 (.08)	7.42 (.10)	1.245 (.004)	.448 (.003)	.845 (.001)	17.17 (.03)	.610 (.001)
	Full	8.63 (.01)	5.52 (.02)	1.194 (.001)	.472 (.003)	.880 (.002)	16.41 (.03)	.622 (.001)
	+T	**8.35 (.08)**	**5.28 (.05)**	**1.192 (.001)**	**.477 (.0004)**	**.886 (.002)**	**16.36 (.02)**	**.623 (.001)**

3.1 Training Dynamics and Checkpoint Selection

Figure 3 shows the representative DeepGlobe phase transition, see Appendix 5.3 for LSUN-Church and BDD100K. The dashed line marks the shared phase-1 initialization. On DeepGlobe, staged-LR continuation improves over ordinary continuation, while the patch-sheaf branches achieve lower FID and better residual calibration. We select checkpoints jointly by FID, residual W_1, mass-ratio proximity to one, and SNR-conditional QCE. This selects the +tail checkpoints at 100k, 390k, and 170k total steps for DeepGlobe, LSUN-Church, and BDD100K, respectively. Although BDD100K attains a marginally lower FID at 160k, 170k provides the best balance across these metrics. Selected results are reported in Table 1, with matched fixed-step controls in Table 2.

3.2 Standard Image Quality and Continuation Controls

Table 1 reports selected checkpoints under the common 50k-sample protocol. Table 2 reports mean ± standard deviation over five independent phase-2 training seeds. Relative to staged LR, +tail reduces mean FID by 11.8%, 12.9%, and 21.4% on DeepGlobe, LSUN-Church, and BDD100K, respectively, while consistently improving KID, CMMD, recall, and coverage. sFID improves on LSUN-Church and BDD100K but not DeepGlobe. Precision decreases on DeepGlobe and LSUN-Church but improves on BDD100K, indicating a dataset-dependent precision–diversity trade-off. Appendix 5.5 reports independent staged-LR no-sheaf, Block I, Blocks I–II, and Blocks I–III branches at matched dataset-specific steps. Each added block reduces FID and KID; the complete objective lowers FID by 7.5%, 14.4%, and 19.1% on DeepGlobe, LSUN-Church, and BDD100K, respectively, relative to staged LR. Appendix 5.6 shows that clean-target matching outperforms zero-residual minimization at every cumulative depth. The matched DeepGlobe control in Appendix 5.7 further separates fixed-map calibration from Block I: Blocks II+III alone reduce FID from 84.083 to 78.635, while adding Block I further reduces it to 77.787, indicating complementary contributions.

3.3 Latent Patch-Space Residual Geometry

Figure 4 shows the timestep-conditioned residual diagnostics on BDD100K; the DeepGlobe and LSUN-Church results are provided in Appendix 5.11. Across all datasets, the no-sheaf controls exhibit high-timestep residual collapse, whereas patch-sheaf training reduces residual W_1 and QCE, keeps the residual-mass ratio closer to one, and improves q95 TailIoU. Table 3 separates the top 5% of patches by pooled Sobel strength ($q_e = 0.95$) from the remaining patches. Relative to the matched staged-LR controls, +tail reduces boundary-region residual W1 by 68.0%, 69.7%, and 71.3% on DeepGlobe, LSUN-Church, and BDD100K, respectively. The improvement also holds in non-boundary regions, showing better calibration near structural transitions and smoother regions.

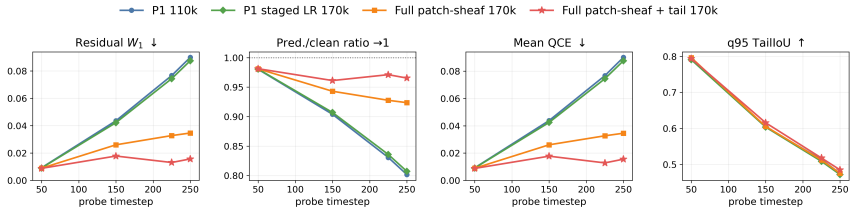

Figure 4: SNR-conditional diagnostics on BDD100K. Patch-sheaf training reduces residual W_1 and QCE, keeps the mass ratio closer to 1, and improves q95 TailIoU, notably at high-noise timesteps.

Table 3: Boundary calibration at $q_e = 0.95$. Red.: W_1 reduction over staged LR.

		Staged LR			Sheaf +tail			
Data	Reg.	$W_1 \downarrow$	QCE↓	Rat.→ 1	$W_1 \downarrow$	QCE↓	Rat.→ 1	Red.↑
DG 100k	B	.0658	.0664	.8944	**.0211**	**.0216**	**.9739**	68.0%
	NB	.0651	.1431	.8228	**.0087**	**.0167**	**.9775**	86.6%
LC 390k	B	.0719	.0976	.8829	**.0218**	**.0377**	**.9649**	69.7%
	NB	.0588	.1032	.8697	**.0109**	**.0207**	**.9758**	81.4%
BDD 170k	B	.0614	.0694	.9067	**.0176**	**.0200**	**.9739**	71.3%
	NB	.0526	.0902	.8809	**.0134**	**.0183**	**.9697**	74.5%

B/NB: boundary/non-boundary; Rat.: predicted-clean/clean residual-mass ratio.

Table 4: Matched image-space W_1 distances (↓); bold marks the best post-transition value.

Data	Model	Edge	Bound.	Grad95	Color	Sat.
DG	P1	.0384	.1492	.2257	3.5295	.0237
	P1C	.0277	.1316	.1891	2.3923	.0173
	SLR	.0284	.1161	.1664	2.0461	.0155
	Full	**.0114**	.0737	.0649	**.7046**	**.0080**
	+T	.0118	**.0726**	**.0630**	.8340	.0085
LC	P1	.0080	.0647	.0900	3.5691	.0229
	P1C	.0104	.0665	.0875	3.9823	.0266
	SLR	**.0042**	.0514	.0750	2.4552	.0165
	Full	.0046	**.0324**	**.0649**	**1.4454**	**.0089**
	+T	.0044	.0385	.0683	1.6593	.0107
BDD	P1	.0024	.0512	.0866	2.6705	.0178
	P1C	.0022	.0546	.0918	2.3716	.0167
	SLR	.0036	.0441	.0893	2.1372	.0126
	Full	.0028	.0095	.0306	1.1023	.0074
	+T	**.0021**	**.0083**	**.0177**	**.9878**	**.0071**

Steps (P1/P2): DG 50/80k, LC 290/390k, BDD 110/170k. P1C: continuation; SLR: staged LR; +T: full sheaf +tail.

Pixel-Space Structure. Table 4 compares generated and real distributions using Wasserstein distances over edge density, boundary contrast, 95th-percentile gradient magnitude, colourfulness, and saturation. On DeepGlobe and BDD100K, both patch-sheaf variants improve all five diagnostics relative to the staged-LR controls. On LSUN-Church, staged LR gives the lowest edge-density distance, while patch-sheaf training substantially improves the remaining boundary, gradient, color, and saturation statistics. These results indicate a structural improvement beyond schedule changes alone. Figure 5 shows visibly greater colour saturation and sharper facade boundaries on LSUN-Church, and clearer, more coherent crosswalk and lane markings on BDD100K for the patch-sheaf variants. Shared seeds facilitate comparison but do not imply pixelwise correspondence.

Cross-Space Micro-Level Geometry. Figure 6 reports patch-boundary orientation coherence at the selected checkpoints. Across all the datasets, the patch-sheaf variants reduce BAE and increase IPGC relative to the no-sheaf controls, with +tail giving the strongest overall results. This indicates more coherent continuation of visible edge geometry across neighbouring patch regions. Figure 7 provides paired fixed-probe examples on LSUN-Church and BDD100K. Relative to staged LR, patch-sheaf training recovers clearer target-aligned traces across the marked patch boundaries, consistent with aggregate BAE and IPGC.

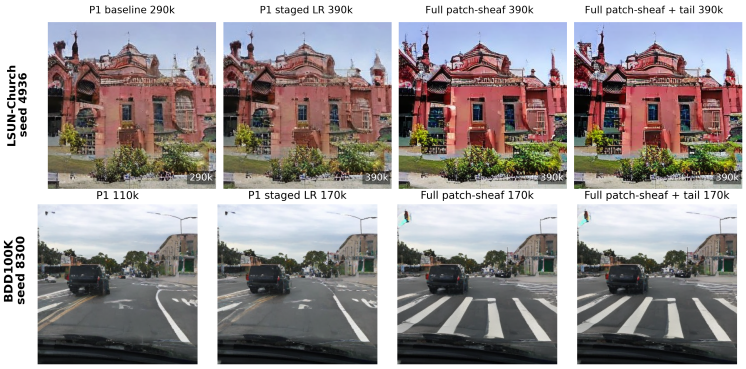

Figure 5: Matched LSUN-Church and BDD100K examples. Sheaf variants improve facade contrast and crosswalk/lane geometry over staged LR.

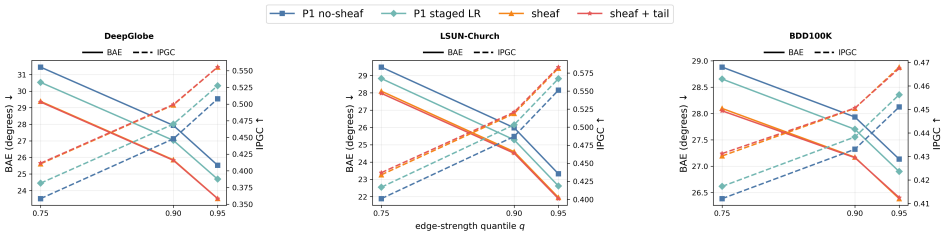

Figure 6: Patch-boundary orientation coherence at selected checkpoints. Lower BAE and higher IPGC indicate more coherent edge continuation across patch boundaries.

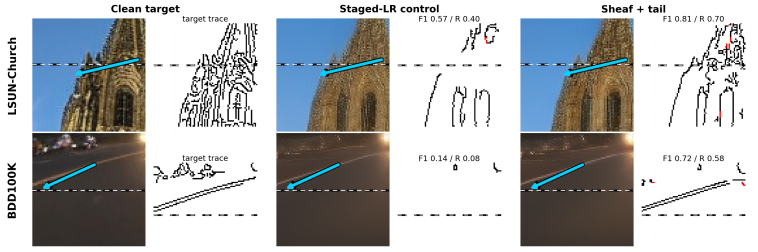

Figure 7: Fixed-probe boundary continuation on LSUN-Church and BDD100K. Dashed lines mark patch boundaries; arrows identify crossing structures; and black/red traces denote matched/unmatched edges. Staged-LR control fails to recover patch connectivity, and transitions.

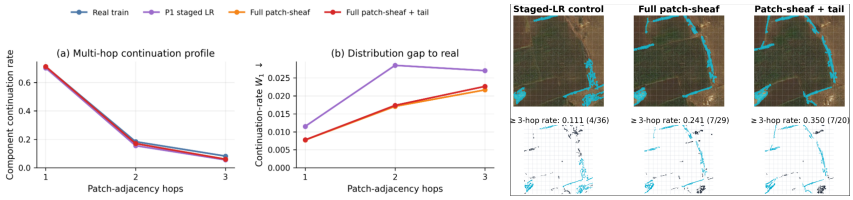

Figure 8: DeepGlobe multi-hop continuation. (Left) Mean rates beyond one hop (a) and W_1 distribution gaps to real images (b). (Right) Multi-hop examples, dark traces are eligible components; cyan traces span at least three patch adjacencies.

Multi-Hop Patch Continuation. BAE and IPGC test a single patch boundary. We additionally define the k-hop rate as the fraction of $q_e = 0.95$ Sobel components spanning at least k adjacencies on the 16×16 patch grid. Figure 8 (left) reports mean rates at two and three hops, where proximity to the real-image curve is desirable; panel (b) reports the corresponding W_1 distribution gap, where lower is better. Relative to staged LR, full/+tail reduce this gap by $40.1/39.1\%$ at two hops and $19.9/16.2\%$ at three hops. Thus, the local objective measurably improves continuation beyond immediate neighbours, although not explicitly model arbitrary scene-wide dependencies. An example in Figure 8 (right) shows that RGB overlays and trace maps mark components spanning at least three adjacencies in cyan; dark traces denote all eligible components. The reported fractions over eligible components illustrate reduced fragmentation, while Figure 8 (left) provides the aggregate evidence.

4 Conclusion

We introduced patch-sheaf residual matching, a phase-2 objective that improves cross-patch structural coherence in predicted-clean diffusion latents while preserving meaningful nonzero clean-latent relation energies. Across the independent training seeds, +tail reduces mean FID relative to matched staged-LR controls by 11.8%, 12.9%, and 21.4% on DeepGlobe Roads, LSUN-Church, and BDD100K, respectively, while consistently improving KID, CMMD, recall, and coverage. Fixed-probe and cumulative-block analyses connect these gains to improved residual calibration, while the zero-residual ablation shows that directly suppressing relation energies causes residual collapse. The DeepGlobe multi-hop analysis further demonstrates improved structural continuation beyond one patch adjacency. These results establish gains beyond additional training or learning-rate scheduling alone. The method propagates structural consistency through overlapping local relations but does not explicitly model arbitrary scene-wide semantic dependencies. Future work can incorporate adaptive long-range relations and evaluate the approach across additional architectures.

REFERENCES

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A ViT backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22669–22679, 2023.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1737–1752. PMLR, 2023.

Cristian Bodnar, Francesco Di Giovanni, Benjamin Paul Chamberlain, Pietro Liò, and Michael M. Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in GNNs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18527–18541, 2022.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11472–11481, 2022.

Justin Michael Curry. *Sheaves, Cosheaves and Applications*. PhD thesis, University of Pennsylvania, 2014. Publicly Accessible Penn Dissertations, No. 1249.

Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 172–181, 2018.

Ishan Deshpande, Ziyu Zhang, and Alexander G. Schwing. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Minsu Gong, Nuri Ryu, Jungseul Ok, and Sunghyun Cho. Edge-aware image manipulation via diffusion models with a novel structure-preservation loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5142–5153, 2026. doi: 10.1109/WACV61042.2026.00499.

Saumya Gupta, Dimitris Samaras, and Chao Chen. TopoDiffusionNet: A topology-aware diffusion model. In *International Conference on Learning Representations*, 2025.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via Min-SNR weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7441–7451, 2023.

Jakob Hansen and Robert Ghrist. Toward a spectral theory of cellular sheaves. *Journal of Applied and Computational Topology*, 3:315–358, 2019. doi: 10.1007/s41468-019-00038-7.

Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9307–9315, 2024.

Yuhan Kang, Hengcan Shi, Hao Liu, Weiying Xie, Leyuan Fang, and Lorenzo Bruzzone. GlobDiffusion: A global consistent diffusion model for large-scale image generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(5):7005–7019, May 2026. doi: 10.1109/TCSVT.2025.3643743.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022.

Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde. Sliced-wasserstein autoencoder: An embarrassingly simple generative model. In *International Conference on Learning Representations*, 2019.

Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. SyncDiffusion: Coherent montage via synchronized joint diffusions. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 2020.

Charlie Nash, Jacob Menick, Sander Dieleman, and Peter Battaglia. Generating images with sparse representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7958–7968. PMLR, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

Aaditya Ramdas, Nicolas Garcia, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

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

Jente Vandersanden, Sascha Holl, Xingchang Huang, and Gurprit Singh. Edge-preserving noise for diffusion models. In *ICLR 2025 Workshop on Deep Generative Models in Machine Learning: Theory, Principle and Efficacy*, 2025.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2636–2645, 2020.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

5 APPENDIX

5.1 AUXILIARY-LOSS IMPLEMENTATION DETAILS

Section statistics. The section-statistics loss is computed from the same pooled patch states used by the sheaf. For each patch scale, it matches the predicted-clean and clean per-channel spatial means and standard deviations using mean absolute error. It additionally matches quantiles of the patch-state norms. The losses are averaged across the active patch scales.

Transport implementations. The one-dimensional patch-distribution loss sorts the transformed incident energies independently for each image and computes their mean absolute difference. For spatial matching, predicted-clean and clean point sets are projected onto shared random unit directions in $\mathbb{R}^3$. Each projection is sorted over patches, and the absolute differences are averaged over samples, patches, and projection directions.

Location-wise residual bands. For patch v, let $\hat{E}_{n,v} = \phi(E_v^{\mathrm{str}}(\hat{x}_{0,n}))$, and let $q_\ell(v)$, $q_h(v)$, and $q_{\mathrm{ext}}(v)$ be the 0.10, 0.90, and 0.99 quantiles, respectively, of the transformed clean energies $\{\phi(E_v^{\mathrm{str}}(x_{0,n}))\}_n$ across the mini-batch. We use

$$\begin{aligned}\mathcal{L}_{\mathrm{band}} = \mathrm{mean}_{n,v}\Big[& w_{\mathrm{under}}\big(q_\ell(v) - \hat{E}_{n,v}\big)_+^2 \\ & + w_{\mathrm{over}}\Big(\big(\hat{E}_{n,v} - q_h(v)\big)_+^2 + \big(\hat{E}_{n,v} - q_{\mathrm{ext}}(v)\big)_+^2\Big)\Big].\end{aligned} \tag{12}$$

Thus, energies within the clean 0.10–0.90 range receive no band penalty; energies above the 0.99 quantile receive both upper-tail penalties. The weights w_{under} and w_{over} control the relative costs of residual collapse and amplification.

Relation weights and restriction maps. Edges have unit weight, and we use $\alpha_{\mathrm{poset}} = \alpha_{\mathrm{cell}} = 0.25$ for the parent and square-cell relations in the reported models. Block I uses adaptive diagonal edge restrictions, bounded around and regularized toward the identity, together with fixed averaging maps for parent and cell relations. Blocks II and III instead recompute identity-based edge residuals and aggregate all incident relation energies using normalization by weighted patch degree.

5.2 TIMESTEP-CONDITIONED TAIL CALIBRATION

The +tail terms use transformed incident energies within $50 \le t \le 250$, intersected with the active SNR window. Calibration uses quantiles $\mathcal{Q} = \{0.50, 0.75, 0.90, 0.99\}$ and timestep-bin edges $(50, 100, 200, 240, 250)$. Let $\mathcal{B}$ contain the nonempty bins, $\delta_{n,q} = Q_{n,q}^{\mathrm{pred}} - Q_{n,q}^{\mathrm{clean}}$, and $T_{n,q}^{\mathrm{pred}}$, $T_{n,q}^{\mathrm{clean}}$ denote the corresponding upper-tail means. We use

$$\mathcal{L}_{\mathrm{calib}} = \mathrm{mean}_{b\in\mathcal{B}}\,\mathrm{mean}_{n\in b,\, q\in\mathcal{Q}}\left[w_{b,q}^-(-\delta_{n,q})_+ + w_{b,q}^+(\delta_{n,q})_+ + w_{b,q}^T(T_{n,q}^{\mathrm{clean}} - T_{n,q}^{\mathrm{pred}})_+\right]. \tag{13}$$

Each tail mean averages the largest $\max\{1, \mathrm{round}((1-q)|V|)\}$ energies. The separate symmetric quantile-loss weight is zero in the reported models. Averaging is performed within each bin and then equally across nonempty bins.

The asymmetric weights adapt using the detached relative error $e_{b,q} = \mathrm{mean}_{n\in b}\,\delta_{n,q}/\max\{|Q_{n,q}^{\mathrm{clean}}|, \epsilon_{\mathrm{stab}}\}$:

$$w_{b,q}^- = \mathrm{clip}(\bar{w}_{b,q}^- e^{-\eta e_{b,q}}, w_{\min}, w_{\max}), \qquad w_{b,q}^+ = \mathrm{clip}(\bar{w}_{b,q}^+ e^{\eta e_{b,q}}, w_{\min}, w_{\max}). \tag{14}$$

We use $\eta = 1$, $w_{\min} = 0.25$, and $w_{\max} = 3$. Clean targets and effective calibration weights are detached.

The mode loss $\mathcal{L}_{\mathrm{mode}}$ uses squared hinge penalties when clean mid- or high-residual patches fall below their respective per-image clean 0.75 or 0.95 quantile thresholds. The localization loss $\mathcal{L}_{\mathrm{loc}}$ combines squared penalties for missed high-residual patches, false positives outside a one-patch-radius neighbourhood of clean high-residual patches, and normalized residual-map mismatch. These terms use interior patches; a separate border penalty suppresses above-threshold energies on the one-patch-wide outer border. Map normalization uses the clean 0.75–0.95 quantile range and clips values to $[-4, 4]$. Localization coefficients differ between $t \le 200$ and $t > 200$. The three losses are combined as in Equation 7.

5.3 Additional Phase-1 to Phase-2 Training Dynamics

Figures 9 and 10 complement the representative DeepGlobe analysis in Figure 3. For both datasets, the phase-2 stopping point is selected by jointly considering FID, residual-distribution W_1, the predicted-clean-to-clean residual mass ratio, and SNR-conditional quantile calibration error. This criterion selects the tail-localized checkpoint at 390k total steps for LSUN-Church and 170k for BDD100K. On BDD100K, although the 160k checkpoint attains a marginally lower FID, the 170k checkpoint provides the strongest overall balance across the four monitored trajectories.

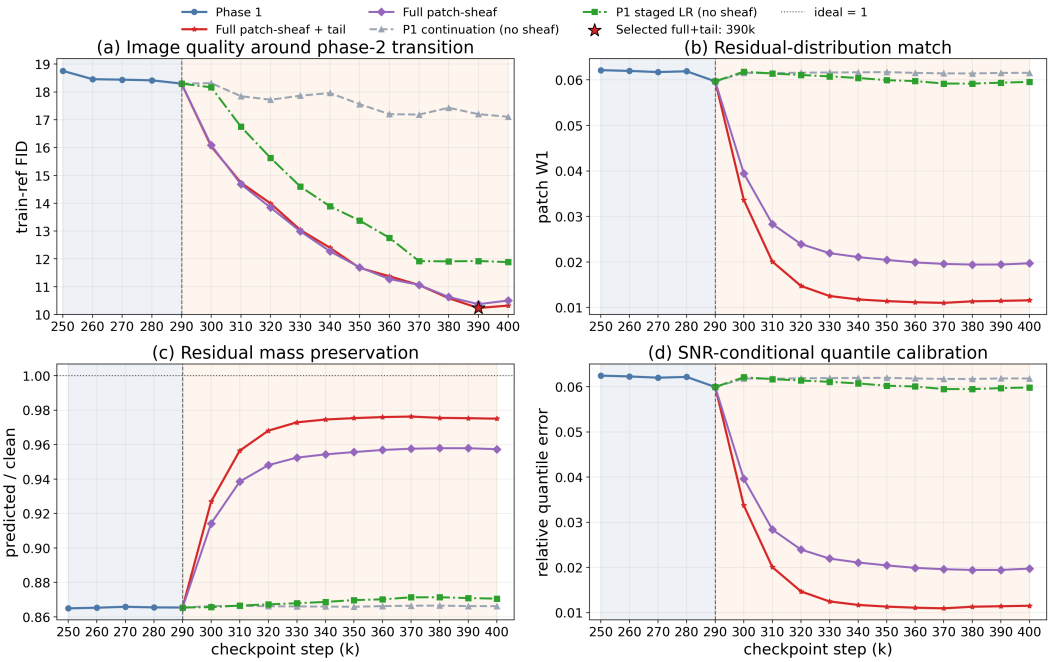

Figure 9: Phase-1 to phase-2 training dynamics on LSUN-Church. The dashed vertical line marks the selected phase-1 checkpoint used to initialize all phase-2 branches. Panel (a) shows the FID trajectories, while panels (b)–(d) show residual-distribution W_1, the predicted-clean-to-clean residual mass ratio, and SNR-conditional quantile calibration error. The joint trajectory analysis selects the tail-localized checkpoint at 390k total steps.

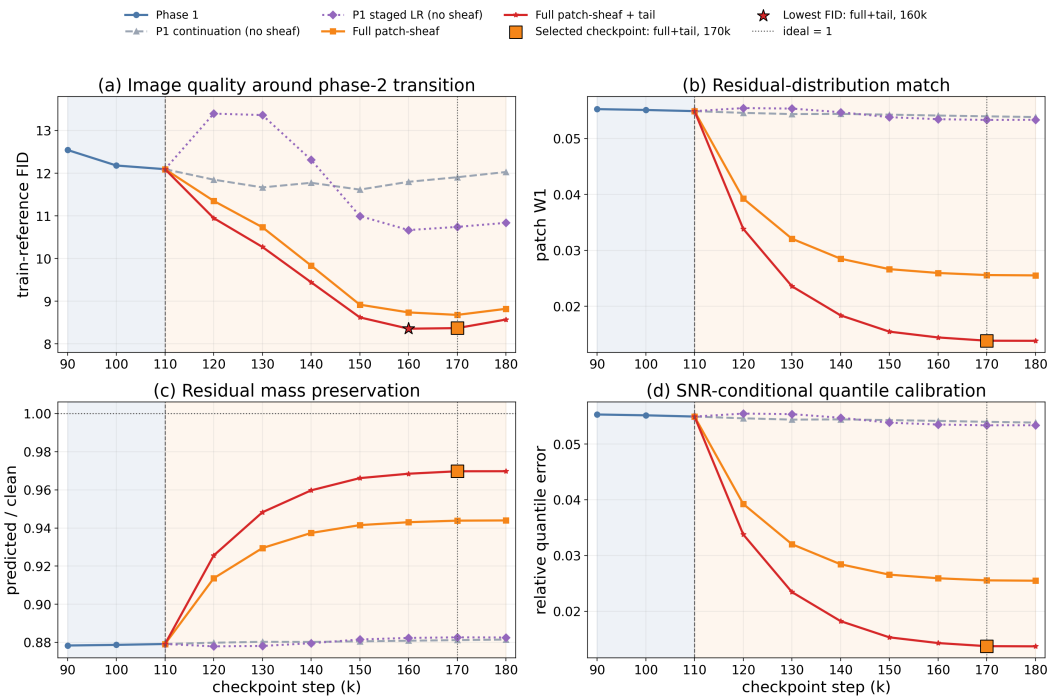

Figure 10: Phase-1 to phase-2 training dynamics on BDD100K. The dashed vertical line marks the selected phase-1 checkpoint used to initialize all phase-2 branches. Panel (a) shows the FID trajectories, while panels (b)–(d) show residual-distribution W_1, the predicted-clean-to-clean residual mass ratio, and SNR-conditional quantile calibration error. Although the tail-localized branch attains a marginally lower FID at 160k, the joint trajectory analysis selects 170k as the overall stopping point.

5.4 Domain-Alignment Ablation on FFHQ

We use FFHQ as a domain-alignment ablation because its comparatively homogeneous and globally constrained facial structure provides less scope for the extended cross-region patch relations to contribute than the complex spatial layouts of DeepGlobe and street-scene datasets. As shown in Figure 11, full patch-sheaf fine-tuning improves both FID and residual calibration relative to ordinary and staged-LR no-sheaf continuations initialized from the same phase-1 checkpoint. The improvement is smaller than on the other datasets, which is consistent with the patch-sheaf inductive bias being especially beneficial for domains containing extended boundaries, connectivity patterns, and heterogeneous neighbouring regions. Nevertheless, the FFHQ results demonstrate that the method remains beneficial even when the data are less strongly aligned with this spatial inductive bias.

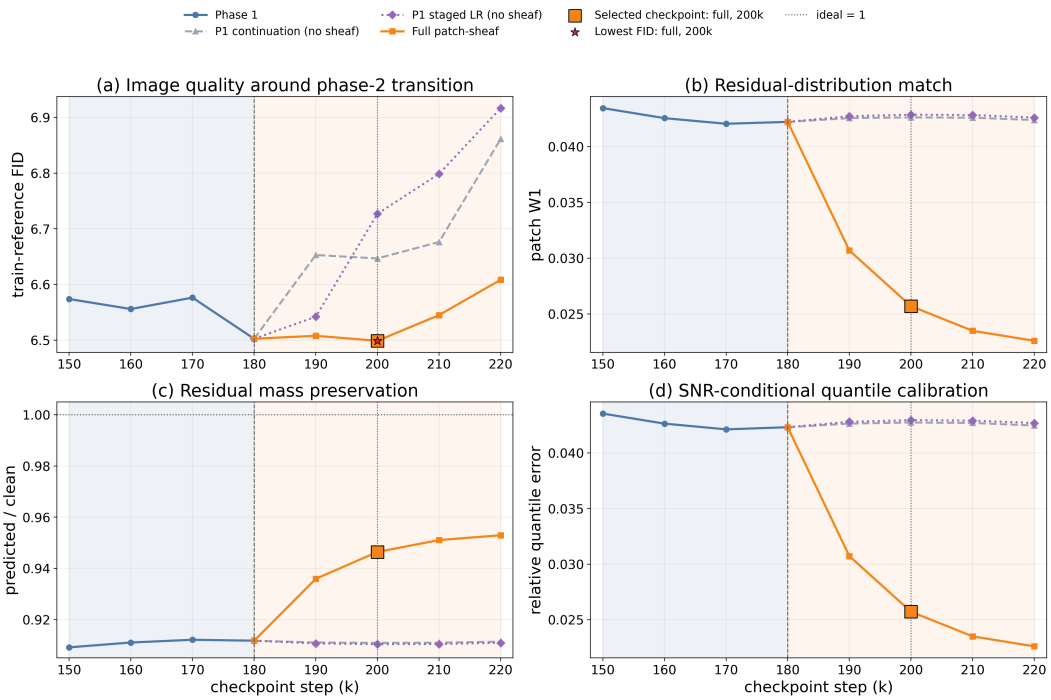

Figure 11: **FFHQ domain-alignment ablation.** All phase-2 branches start from the same 180k checkpoint. Full patch-sheaf training slightly improves FID and more substantially improves residual calibration relative to ordinary and staged-LR no-sheaf continuations.

5.5 Incremental Analysis of the Patch-Sheaf Objective

We evaluate the three objective blocks cumulatively using independently trained phase-2 branches. Within each dataset, every branch starts from the same phase-1 checkpoint, follows the same staged-LR schedule, and is evaluated at the matched total step used in Table 2: 100k for DeepGlobe, 390k for LSUN-Church, and 170k for BDD100K. The enabled blocks remain active throughout training; they are not introduced sequentially within one run. This controls for initialization, optimization schedule, and training duration, without per-variant checkpoint selection.

Table 5 shows monotonic FID and KID improvements as the three blocks are added. Relative to staged LR, Block I reduces FID by 1.8%, 6.6%, and 5.7% on DeepGlobe, LSUN-Church, and BDD100K, respectively. Block II provides further reductions of 4.7%, 6.4%, and 9.1%, while Block III contributes another 1.2%, 2.1%, and 5.5%. Overall, Blocks I–III reduce FID by 7.5%, 14.4%, and 19.1% relative to staged LR. Despite small dataset-dependent fluctuations in secondary metrics, every added block improves FID and KID, supporting the complete objective.

5.6 Zero-Residual versus Clean-Target Matching

We directly test the motivation for clean-target residual matching on DeepGlobe. For $A \in \{E, P, C\}$, the zero-residual variant minimizes the transformed predicted relation energy, whereas the proposed objective matches it to the corresponding clean relation energy:

$$\begin{aligned} \mathcal{L}_A^{\text{zero}} &= \text{mean}_i\, \phi(\varepsilon_{A,i}(\hat{x}_0))\,, \\ \mathcal{L}_A^{\text{match}} &= \text{mean}_i\, [\phi(\varepsilon_{A,i}(\hat{x}_0)) - \phi(\varepsilon_{A,i}(x_0))]^2\,, \end{aligned} \tag{15}$$

where $\phi(s) = \sqrt{\max(s, \epsilon)}$. Thus, $\mathcal{L}_A^{\text{zero}}$ directly suppresses relation energies, whereas $\mathcal{L}_A^{\text{match}}$ preserves the nonzero incompatibilities observed in clean latents. We compare the staged-LR no-sheaf control with zero-residual and clean-target variants at three cumulative objective depths: Block I, Blocks I–II, and Blocks I–III. Every branch starts from the same 50k phase-1 checkpoint, follows the same staged learning-rate schedule, and is evaluated at 100k total steps. At each depth, only

Table 5: Cumulative block ablation using 50k samples at matched checkpoints. Blocks are added cumulatively. KID is multiplied by 10^3. Bold marks the best result per dataset.

Dataset	Step	Variant	I	II	III	FID ↓	KID ↓	CMMD ↓	R ↑	Cov. ↑	sFID ↓
DeepGlobe	100k	P1 + staged LR	✗	✗	✗	84.083	68.00	1.8276	0.169	0.438	41.659
		Block I	✓	✗	✗	82.589	65.87	1.8291	0.185	0.424	**41.621**
		Blocks I–II	✓	✓	✗	78.740	61.12	**1.7820**	**0.190**	0.438	42.575
		Blocks I–III	✓	✓	✓	**77.787**	**59.79**	1.7882	0.187	**0.446**	42.516
LSUN-Church	390k	P1 + staged LR	✗	✗	✗	12.080	8.84	2.2999	0.408	0.741	27.664
		Block I	✓	✗	✗	11.279	8.15	2.2283	0.433	0.744	26.548
		Blocks I–II	✓	✓	✗	10.559	7.44	2.1925	0.445	**0.746**	25.284
		Blocks I–III	✓	✓	✓	**10.337**	**7.13**	**2.1750**	**0.446**	**0.746**	**25.253**
BDD100K	170k	P1 + staged LR	✗	✗	✗	10.642	7.43	1.2445	0.449	0.845	17.156
		Block I	✓	✗	✗	10.033	6.84	1.2144	0.456	0.858	16.797
		Blocks I–II	✓	✓	✗	9.120	5.95	1.2006	0.465	0.876	16.454
		Blocks I–III	✓	✓	✓	**8.614**	**5.54**	**1.1926**	**0.470**	**0.880**	**16.427**

Table 6: DeepGlobe zero-residual ablation at 100k total steps. Mean metrics average $t \in \{150, 225, 250\}$. At each cumulative depth, only the Block-I target is changed; Blocks II and III retain their original objectives. Lower W_1 is better, and the residual-mass ratio should approach one. Bold marks the better target at each depth.

Depth	Block-I target	Mean W_1 ↓	Mean ratio → 1	$W_1, t = 250$ ↓	Ratio, $t = 250 \to 1$
No sheaf	Staged-LR control	0.0832	0.782	0.1064	0.719
Block I	Zero residual	0.0999	0.737	0.1299	0.657
	Clean target	**0.0665**	**0.829**	**0.0822**	**0.788**
Blocks I–II	Zero residual	0.0536	0.862	0.0638	0.835
	Clean target	**0.0274**	**0.934**	**0.0272**	**0.937**
Blocks I–III	Zero residual	0.0451	0.885	0.0519	0.868
	Clean target	**0.0223**	**0.948**	**0.0208**	**0.954**

the Block-I relation target is changed. Blocks II and III, when enabled, retain their original clean-distribution and spatial matching objectives. Adaptive restrictions, relation weights, section statistics, the SNR gate, and all other settings are held fixed. Fixed-probe diagnostics use identical clean images, timesteps, and noise realizations.

At Block I, zero-residual minimization performs worse than the staged-LR control, increasing residual W_1 and moving the residual-mass ratio farther from one. Clean-target matching instead reduces mean W_1 by 33.4% relative to the zero-residual variant and better preserves residual mass. Blocks II and III partially counteract zero-residual suppression through their distributional and spatial objectives; nevertheless, clean-target matching remains consistently superior at every cumulative depth. Its mean-W_1 reduction over zero-residual minimization increases to 48.9% with Blocks I–II and 50.6% with Blocks I–III, while the corresponding mass ratios improve from 0.862 to 0.934 and from 0.885 to 0.948. These results directly support preserving clean, nonzero relation energies rather than forcing patch incompatibilities to vanish. These results show that zero-residual minimization suppresses data-like residual variation, a failure mode consistent with latent oversmoothing. Clean-target matching avoids this collapse by preserving the nonzero relation energies observed in clean latents.

5.7 Relation-Level Sheaf Geometry versus Residual Calibration

The cumulative ablation measures the benefit of adding each block, but does not separate relation-level sheaf geometry from fixed-map residual calibration. We therefore train an additional *Blocks II+III only* DeepGlobe branch. It retains the incident-energy distribution, location-band, and spatial sliced-Wasserstein objectives, while disabling Block I clean-target relation matching, adaptive edge restrictions, restriction regularization, and section statistics. All branches start from the same 50k phase-1 checkpoint, use the same staged learning-rate schedule, and are evaluated at 100k total steps with 50k generated samples.

Blocks II+III alone reduce FID by 6.5% relative to staged LR, showing that fixed-map distributional and spatial calibration provide a substantial gain. Adding Block I further reduces FID by 1.1% and

Table 7: DeepGlobe geometry attribution at the matched 100k checkpoint using 50k generated samples. KID is multiplied by 10^3.

Method	FID↓	KID↓	CMMD↓	Rec.↑	Cov.↑	sFID↓	Prec.↑
Staged LR	84.083	68.00	1.8276	0.169	0.438	**41.659**	**0.349**
Blocks II+III only	78.635	60.60	**1.7710**	0.180	**0.464**	42.859	0.339
Blocks I–III	**77.787**	**59.79**	1.7882	**0.187**	0.446	42.516	0.330

KID by 1.3%, while raising recall from 0.180 to 0.187. CMMD and coverage instead favor Blocks II+III alone. Thus, Block I's clean-target relation matching, adaptive edge maps, restriction regularization, and section statistics make an additional but smaller contribution, with complementary metric trade-offs.

5.8 Extra qualitative results

Paired structural recovery. Figure 12 visualizes the micro-level effect on a fixed diagnostic probe. All methods receive the same clean target x_0, timestep $t = 250$, and noise realization ϵ. The sheaf models recover more target-aligned structural edges than the staged-LR control, indicating better preservation of local structures in the decoded image space.

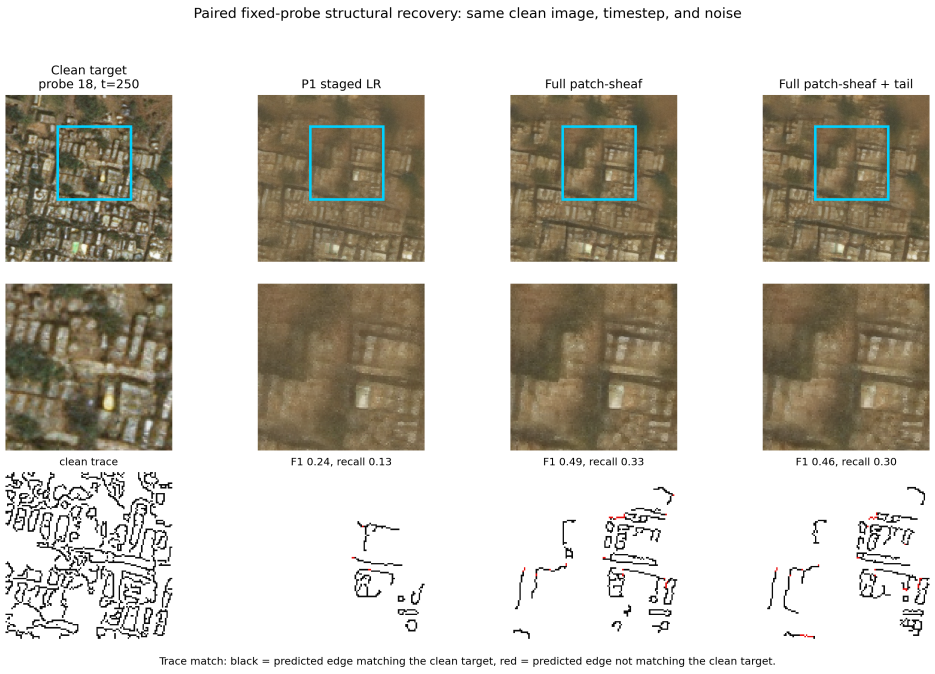

Figure 12: Paired structural recovery from the same noisy latent. Black edges match the clean target; red edges are predicted but not target-aligned. Patch-sheaf models recover more target-aligned structure than the staged-LR control.

Additional fixed-seed generation. Figure 13 provides the corresponding DeepGlobe comparison omitted from the main qualitative figure.

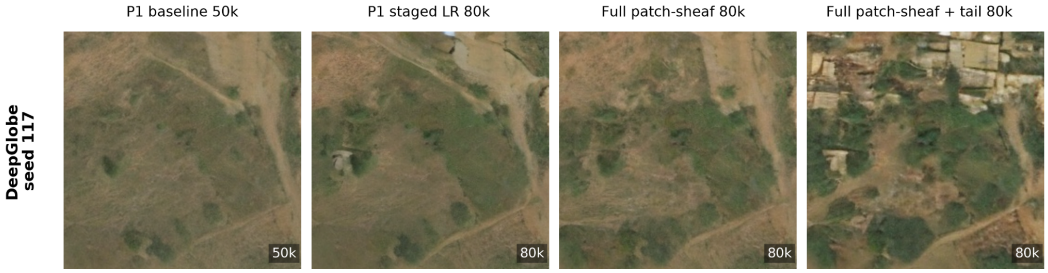

Figure 13: Matched DeepGlobe fixed-seed example. The sheaf variants show clearer land-cover separation and stronger local contrast than staged LR.

5.9 Sensitivity Analysis on BDD100K

We perform a one-factor-at-a-time sensitivity analysis of the phase-2 training configuration on BDD100K. We vary the learning-rate-shock interval, active signal-to-noise ratio (SNR) window, and one primary weight from each auxiliary block. Patch resolution is examined separately in Appendix 5.10, because it changes the spatial discretization of the patch complex and exhibits a stronger, dataset-dependent effect.

A learning-rate shock starts a new training segment with a reset optimizer state and the learning rate prescribed by the phase-2 schedule; the reference configuration uses 10k-step intervals. All variants start from the same selected phase-1 checkpoint at 110k and are evaluated at 170k, giving a matched 60k-step phase-2 budget. Within each group, only the indicated parameter varies; all remaining settings are fixed to the reference full patch-sheaf configuration. No per-variant checkpoint selection is performed.

Table 8: One-factor-at-a-time optimization and objective sensitivity on BDD100K using 50k generated samples. All variants start at 110k and are evaluated at 170k. The dagger marks the fixed reference configuration; only the indicated parameter varies. KID is multiplied by 10^3, and bold marks the best value within each group.

Group	Parameter	Setting	FID ↓	KID ↓	CMMD ↓	Rec. ↑	Cov. ↑
Optimization	LR-shock interval	None	9.511	6.464	1.2146	.4700	.8680
		5k	8.618	5.540	1.1944	**.4738**	.8799
		10k †	**8.614**	**5.538**	**1.1926**	.4702	**.8801**
		20k	9.528	6.494	1.2202	.4699	.8676
Denoising	Active SNR window	$[1, 5]$	9.250	6.118	1.2190	.4640	.8710
		$[0.5, 10]$ †	8.614	5.538	1.1926	.4702	.8801
		$[0.25, 20]$	**8.363**	**5.235**	**1.1744**	**.4755**	**.8838**
Block I	λ_{sheaf}	0.005	8.633	5.548	1.1961	.4724	.8807
		0.010 †	8.614	5.538	**1.1926**	.4702	.8801
		0.020	**8.558**	**5.495**	1.1936	**.4749**	**.8853**
Block II	$\lambda_{\mathrm{patchdist}}$	0.005	8.966	5.872	1.2003	.4715	.8712
		0.010 †	8.614	5.538	1.1926	.4702	.8801
		0.020	**8.176**	**5.118**	**1.1885**	**.4780**	**.8913**
Block III	$\lambda_{\mathrm{jointpatch}}$	0.001	8.775	5.690	1.1970	.4717	.8785
		0.002 †	8.614	5.538	1.1926	.4702	.8801
		0.004	**8.334**	**5.274**	**1.1875**	**.4766**	**.8850**

Table 8 shows that the patch-sheaf gains persist across the tested optimization, denoising, and objective-weight settings. The 5k and 10k LR-shock intervals perform nearly identically, while the wider SNR window and larger Block-II and Block-III weights provide moderate additional gains. Across all tested settings, FID remains below the matched staged-LR control reported in Table 2,

Table 9: Cross-dataset patch-resolution analysis using 50k generated samples. Within each dataset, only the patch size varies; all branches use matched checkpoints and otherwise identical settings. The dagger marks the shared reference configuration used in the main experiments. KID is multiplied by 10^3; bold marks the best result per dataset.

Dataset (step)	Patch size	FID ↓	KID ↓	CMMD ↓	Rec. ↑	Cov. ↑
DeepGlobe (100k)	1	**74.967**	**55.73**	1.9930	.131	**.490**
	$2^\dagger$	77.787	59.79	**1.7882**	**.187**	.446
	4	83.610	66.51	1.8561	.183	.418
LSUN-Church (390k)	1	13.447	10.54	**2.1161**	.401	.654
	$2^\dagger$	**10.337**	**7.13**	2.1750	**.446**	**.746**
	4	11.934	8.76	2.2405	.424	.732
BDD100K (170k)	1	**6.828**	**4.08**	**1.1676**	**.483**	**.909**
	$2^\dagger$	8.614	5.54	1.1926	.470	.880
	4	10.549	7.44	1.2288	.459	.848

indicating robustness around the fixed reference configuration. The dagger denotes this reference setting rather than a dataset-specific optimum. Because parameters are varied independently, their improvements should not be interpreted as additive.

Patch resolution produces a stronger and less uniform effect than these scalar controls. We therefore examine it separately across all three datasets in Appendix 5.10.

5.10 Patch Resolution as a Dataset-Dependent Design Choice

Unlike the scalar controls above, patch size changes the spatial granularity of the patch complex and its residual relations. Because the initial BDD100K analysis showed a comparatively strong resolution effect, we repeat the matched patch-size comparison on all three datasets. Within each dataset, only the patch size varies; training budgets, evaluation checkpoints, and all other settings remain fixed.

Table 9 reveals a dataset-dependent scale preference rather than a universal optimum. Size 2 gives the strongest overall LSUN-Church results and a balanced DeepGlobe profile. Size 1 improves all reported BDD100K metrics and improves DeepGlobe FID, KID, and coverage, but increases its CMMD and recall. Size 4 is never optimal. Thus, size 2 remains a reasonable shared reference configuration, while finer discretization provides substantial headroom on dense street scenes.

5.11 Additional SNR-Conditional Residual Diagnostics

Figure 14 complements the BDD100K results in Figure 4 with DeepGlobe and LSUN-Church. All methods are evaluated on the same fixed probe of clean images, timesteps, and noise realizations. On both datasets, the no-sheaf controls show increasing residual mismatch at higher-noise timesteps, whereas patch-sheaf training reduces residual W_1 and QCE, keeps the predicted-clean/clean mass ratio closer to one, and improves high-residual localization. The $+$tail variant provides the strongest overall residual calibration.

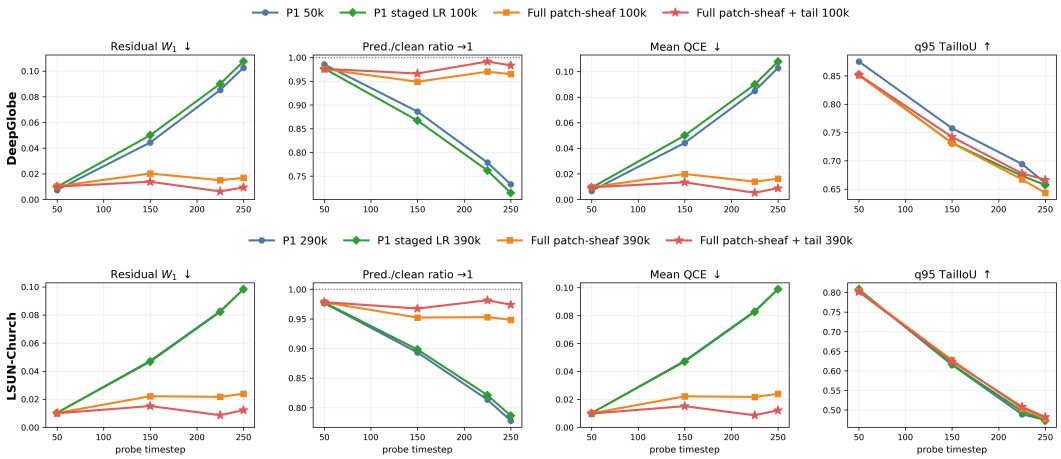

Figure 14: Fixed-probe SNR diagnostics on DeepGlobe (top) and LSUN-Church (bottom). Lower W_1/QCE, ratio near one, and higher q95 TailIoU are better.