CONFORMAL SELECTIVE GENERATION: REGIONAL DELIVERY WITH PROXY RISK CONTROL FOR IMAGE INPAINTING

Anonymous authors
Working draft

ABSTRACT

Image inpainting produces plausible completions, yet deployment requires deciding which generated regions to show. We introduce Conformal Selective Generation (CSG), a framework for regional selective release of a completion designated before scoring. CSG separates the deliverable from auxiliary samples used to estimate uncertainty. It applies a fixed regional score and calibrates a RELEASE/ABSTAIN threshold through conformal risk control across sources. A source loss with a fixed denominator measures released proxy errors and preserves dependence among regions and repeated instances from the same source. Under a frozen protocol and source exchangeability, calibration controls marginal expected proxy risk for any score fixed during development that induces nested release sets. Our primary score measures DINOv2 disagreement between the designated output and the auxiliary centroid. We also evaluate DINO, CLIP, LPIPS, and scores adapted to regional release under the same protocol. On a benchmark of natural images, the primary instantiation achieves 65.3% mean region coverage at mean proxy risk 0.098 across 200 paired source re-splits. Alternative scores yield different risk–coverage tradeoffs. Experiments across four restorers, auxiliary bank sizes, human judgments, perturbations, and mask geometries characterize utility and applicability limits. CSG provides a protocol for partial delivery under an explicit proxy risk budget, with the regional score as a separate modeling choice.

1 INTRODUCTION

Image inpainting can produce coherent scenes with details unsupported by the observation. When a completion serves as a record, this ambiguity creates a delivery problem: rejecting it discards useful content, while displaying it gives uncertain details the same status as observed pixels. Regional delivery preserves observations and selectively withholds generated regions (Figure 1).

Generative restorers recover missing image content (Chung et al., 2023; Wang et al., 2023; Martin et al., 2025); uncertainty methods provide calibrated intervals, image losses, or correlated directions (Horwitz & Hoshen, 2022; Teneggi et al., 2023; Belhasin et al., 2024; Adame et al., 2025). TAIR and ImageDoctor address specific fidelity requirements (Min et al., 2026; Guo et al., 2026). Kutiel et al. already calibrate spatial masks for existing image predictions, including completion (Kutiel et al., 2022). CSG combines discrete regional release, separate auxiliary scoring samples, and proxy error burden averaged within source identity. A source is an underlying image; an instance pairs that image with a mask. Instances from the same source may be dependent.

Selective prediction trades coverage for error (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017), with extensions to deferral and shift (Nguyen et al., 2025; Joren et al., 2024; Narasimhan et al., 2024). CRC calibrates bounded monotone losses (Angelopoulos et al., 2024), with extensions to nonexchangeable settings (Farinhas et al., 2024). Related conformal prediction methods incorporate downstream utility while retaining coverage guarantees (Cortes-Gomez et al., 2025). Conformal Language Modeling, SCOPE-Gen, Sample-then-Identify, and SAFER calibrate generation or selection (Quach et al., 2024; Kladny et al., 2025; Wang et al., 2025; 2026); CREDO assesses a prescribed decision (Zhou et al., 2026). Regional delivery requires aligning the released object, source unit, and error denominator (Appendix A.1).

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.

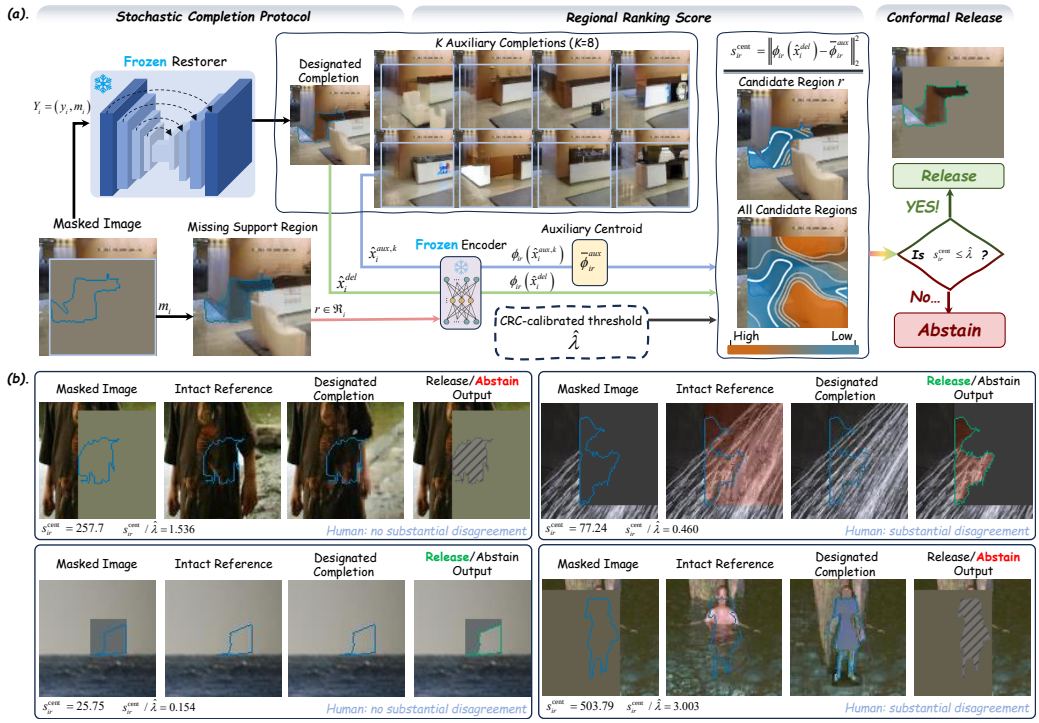

Figure 1: **CSG separates the released completion from its scoring samples.** One output is designated before scoring; K disjoint auxiliary completions provide regional disagreement. Calibration across sources converts this score into a RELEASE/ABSTAIN mask on missing pixels. Intact references supply offline proxy labels; test references are used only for evaluation. Here $K = 8$ counts auxiliaries; ϕ and s_{ir}^{cent} denote regional features and centroid scores (Section 2.2). The examples show partial delivery of the designated output, with observed pixels preserved.

Two difficulties shape this problem. First, samples can agree because they share an error; diversity alone needs careful interpretation (Cohen et al., 2024; Xu et al., 2025; Zheng et al., 2025). Cross-model disagreement has been studied in language models (Hamidieh et al., 2026); CSG samples one frozen restorer and fixes the deliverable before scoring, avoiding content-based changes to the assessed object. Second, regions share content and errors within sources. Calibration must retain this dependence and match the source-level sampling unit encountered at deployment. CSG connects regional release to a proxy risk budget, treating the score as a separate modeling choice.

We make three contributions. First, we study regional delivery when the completion eligible for release is fixed before uncertainty scoring. The joint delivery and loss protocol specifies the deployed object and regional action and excludes content-based output selection; CRC does not generally require predesignation.

Second, we calibrate a nested regional release family from any score fixed during development, using fixed denominators to keep source losses bounded and monotone in the threshold, as CRC requires. Aggregation within instances and then sources preserves dependence among regions and repeated instances. Existing conformal risk control gives a marginal expected proxy risk guarantee under source exchangeability.

Third, we instantiate the framework with DINO centroid disagreement and compare alternative scores under the same protocol. Paired comparisons, four restorers, studies of auxiliary banks, human audits, and geometry and shift tests characterize utility and applicability limits, distinguishing score behavior from calibration properties.

2 METHODOLOGY

As shown in Figure 1, CSG first fixes the completion and candidate regions, then constructs a score, a proxy loss, and a calibrated release rule.

2.1 FIXED COMPLETION AND CANDIDATE REGIONS

For masked input $Y_i = (y_i, m_i)$, let $M_i = \{p : m_i(p) = 1\}$ be the missing support. A frozen stochastic restorer f produces one designated output and K auxiliary completions, as formulated in Equation (1):

$$\hat{x}_i^{\mathrm{del}} = f(Y_i; z_i^{\mathrm{del}}), \qquad \mathcal{X}_i^{\mathrm{aux},K} = \{\hat{x}_i^{\mathrm{aux},k} = f(Y_i; z_{ik}^{\mathrm{aux}})\}_{k=1}^{K}, \tag{1}$$

The z variables specify stochastic states or cached run identities. Designated and auxiliary roles are disjoint and fixed before scoring; independence among outputs is not assumed. The designated completion stays fixed as K changes, without content-based selection. Each uncached instance requires $K + 1$ restorer calls; shared Stable Diffusion inpainting (SD-inpaint) regions add one SD-inpaint call for other restorers. Predesignation defines our deployment contract. CRC can also accommodate a selector fixed during development and independent of the threshold when the resulting source losses remain bounded, nested, and exchangeable.

A frozen regionizer partitions available image information into disjoint base regions B. We retain regions satisfying $|B \cap M_i|/|B| \geq \rho$, where ρ is fixed during development. Connected components of each retained intersection form the candidate set $\mathcal{R}_i$, with $R_i = |\mathcal{R}_i|$. Observed pixels are excluded; missing pixels outside this candidate support receive ABSTAIN. The shared-map comparison reuses the designated SD-inpaint partition to fix candidate regions across restorers (Appendix B.5); a native-map ablation rebuilds each restorer's partition (Appendix C.2.3).

2.2 REGIONAL SCORING FUNCTIONS

CSG uses a scalar regional score $s_{ir} = S(Y_i, \hat{x}_i^{\mathrm{del}}, \mathcal{X}_i^{\mathrm{aux},K}, r)$, with S fixed using development data before calibration. Lower scores indicate greater release suitability. Threshold λ induces the nested family $\mathcal{A}_i(\lambda) = \{r \in \mathcal{R}_i : s_{ir} \leq \lambda\}$. Calibration in Sections 2.3–2.4 depends on this nested family and the loss for each source, but not on the analytic form of S. Scores can share the released object, candidate regions, proxy labels, and calibration rule.

Primary instantiation. Our primary experiments use disagreement between the designated output and the auxiliary centroid in DINOv2 feature space. Let $\phi_{ir}(x)$ denote the feature for region r of image x. We interpolate the dense token map to image resolution before regional mean pooling. The backbone uses LayerNorm. We do not apply L2 normalization after pooling, retaining differences in pooled-vector magnitude. The centroid and primary score are

$$\bar{\phi}_{ir}^{\mathrm{aux}} = \frac{1}{K}\sum_{k=1}^{K} \phi_{ir}(\hat{x}_i^{\mathrm{aux},k}), \qquad s_{ir}^{\mathrm{cent}} = \left\|\phi_{ir}(\hat{x}_i^{\mathrm{del}}) - \bar{\phi}_{ir}^{\mathrm{aux}}\right\|_2^2. \tag{2}$$

This score measures how atypical the designated completion is relative to the auxiliary center. Because the completion is fixed before scoring, s_{ir}^{cent} directly evaluates the object eligible for release.

Alternative score constructions. The framework does not require centroid distance. Auxiliary spread measures variation within the auxiliary bank; mean squared distance from the designated output to each auxiliary combines designated deviation and auxiliary dispersion:

$$s_{ir}^{\mathrm{spr}} = \frac{1}{K}\sum_{k=1}^{K} \left\|\phi_{ir}(\hat{x}_i^{\mathrm{aux},k}) - \bar{\phi}_{ir}^{\mathrm{aux}}\right\|_2^2,$$

$$s_{ir}^{\mathrm{pair}} = \frac{1}{K}\sum_{k=1}^{K} \left\|\phi_{ir}(\hat{x}_i^{\mathrm{del}}) - \phi_{ir}(\hat{x}_i^{\mathrm{aux},k})\right\|_2^2 = s_{ir}^{\mathrm{cent}} + s_{ir}^{\mathrm{spr}}.$$

Appendix A.2 proves the identity. With auxiliary samples and regions fixed, auxiliary-only scores are invariant to changes confined to the designated completion. Scores that include that completion

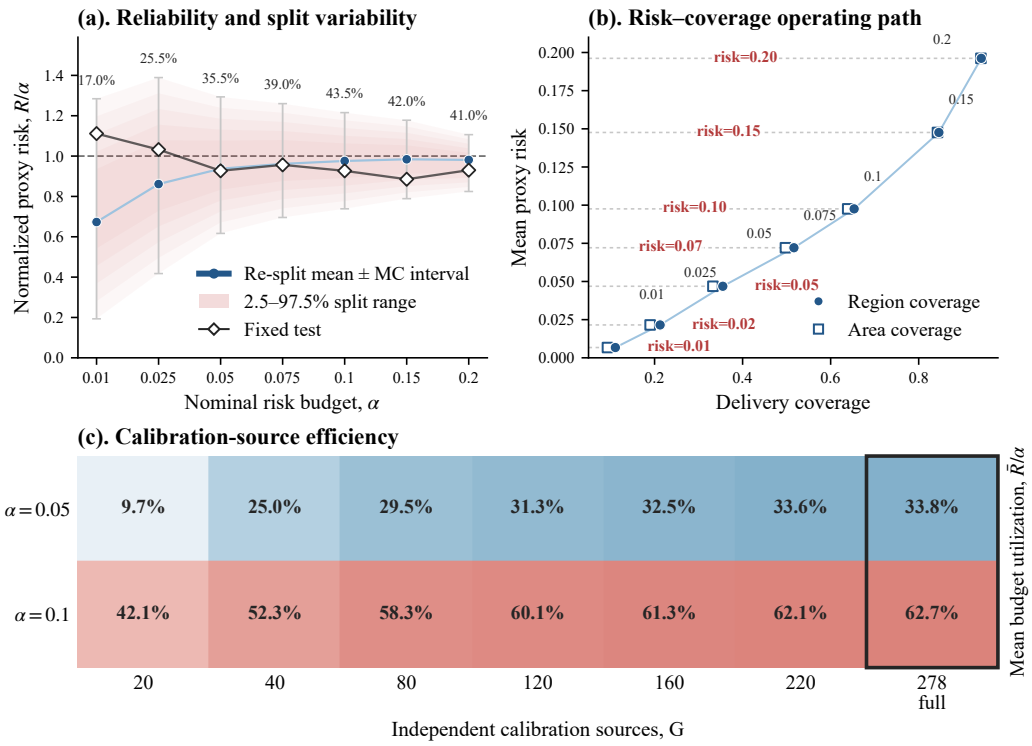

Figure 2: **Calibration behavior of the primary instantiation on SD-natural.** (a) Proxy risk averaged over sources, split percentile ranges, and fixed-test risk; percentages denote split exceedances. (b) Risk versus region/area coverage; horizontal reference lines indicate nominal risk budgets α. (c) Region coverage versus calibration-source count, shaded by R/α. Panel (a) uses the shared re-splits in Section 3.2. CRC controls marginal expected loss.

can respond; centroid and pair mean have identical raw score increments, although their calibrated decisions can differ. Section 3.4 tests this distinction. The DINOv2 centroid score supplies the main analyses of operating points; other DINO, CLIP, LPIPS, and scores adapted to the task distinguish scoring behavior from properties of the calibration protocol.

2.3 Reference-Relative Proxy Loss

For any frozen regional score s_{ir} from Section 2.2, calibration uses an intact reference $x_i^\star$ to define a proxy error target. A separate CLIP encoder supplies regional features ψ_{ir} by interpolating patch tokens and mean-pooling within regions (Appendix B). A development-fixed threshold τ defines proxy errors, aggregated over candidate regions and within sources in Equation (3):

$$\begin{aligned} \tilde{h}_{ir} &= \mathbf{1}\left\{1 - \frac{\langle \psi_{ir}(\hat{x}_i^{\mathrm{del}}), \psi_{ir}(x_i^\star)\rangle}{\|\psi_{ir}(\hat{x}_i^{\mathrm{del}})\|_2 \|\psi_{ir}(x_i^\star)\|_2} > \tau\right\}, \\ \ell_i(\lambda) &= \begin{cases} R_i^{-1} \sum_{r \in \mathcal{R}_i} \tilde{h}_{ir} \mathbf{1}\{s_{ir} \le \lambda\}, & R_i > 0, \\ 0, & R_i = 0, \end{cases} \qquad L_g(\lambda) = \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} \ell_i(\lambda), \end{aligned} \tag{3}$$

Here $\tilde{h}_{ir}$ indicates proxy error, ℓ_i is the instance loss, and L_g averages instances $\mathcal{I}_g$ from source g. Fixed candidate denominators make $L_g \in [0, 1]$ nondecreasing in λ. Empty candidates have zero loss and coverage; abstention contributes zero loss. Each erroneous released region contributes equally relative to the candidate support. This loss budgets exposed errors without guaranteeing release precision; we report conditional error separately. Fixed-denominator area-weighted losses also admit this calibration (Appendix I).

2.4 CONFORMAL RELEASE

Development data fix the rule for designating the output, protocol for sampling auxiliaries, regionizer, regional scoring rule S, score representation, proxy target, K, τ, and finite threshold grid Λ. Calibration selects only the release threshold from G source groups. We select the largest threshold satisfying the CRC criterion at risk budget α, as formulated in Equation (4):

$$\widehat{B}_G(\lambda) = \frac{\sum_{g=1}^{G} L_g(\lambda) + 1}{G+1}, \qquad \hat{\lambda} = \max\{\lambda \in \Lambda : \widehat{B}_G(\lambda) \le \alpha\}, \tag{4}$$

where $\widehat{B}_G$ corrects empirical source loss for the finite calibration sample, using the unit upper bound on each source loss, and $\hat{\lambda}$ is the selected threshold. An empty feasible set returns $\hat{\lambda} = \perp$, meaning complete abstention.

Proposition 1 (Marginal proxy risk control across sources)**.** *Let s_{ir} be any regional scoring rule fixed during development with nested thresholded release sets: $\{r \in \mathcal{R}_i : s_{ir} \le \lambda_1\} \subseteq \{r \in \mathcal{R}_i : s_{ir} \le \lambda_2\}$ whenever $\lambda_1 \le \lambda_2$, for every instance i. Suppose the masking procedure, rule for designating the output, rule for sampling auxiliaries, regionizer, proxy target, and threshold grid are fixed before calibration. If calibration source groups and a future source group are exchangeable under this protocol, the threshold selected by Equation* (4) *satisfies*

$$\mathbb{E}_{\mathcal{D}_{\mathrm{cal}},\, \mathcal{G}_{G+1},\, Z}[L_{G+1}(\hat{\lambda})] \le \alpha. \tag{5}$$

The expectation averages over calibration groups $\mathcal{D}_{\mathrm{cal}}$, a future source group $\mathcal{G}_{G+1}$, and protocol randomness Z; $L_{G+1}(\hat{\lambda})$ is that source's fixed-denominator proxy loss.

The result applies CRC (Angelopoulos et al., 2024) to bounded nested source losses. Regions and repeated instances within a source may be dependent. The guarantee is marginal over future sources and applies to the proxy loss. It does not guarantee error control for individual splits or subgroups, conditional on release, or under human judgments. Appendix A.3 gives the proof and edge cases.

At deployment, CSG releases designated pixels in region r when $s_{ir} \le \hat{\lambda}$. Other missing pixels receive a visible abstention pattern; observed pixels remain unchanged. Algorithm 2 in Appendix A.4 acts on one instance; the guarantee concerns sources averaged over their instance sets (Appendix J.1). For the primary instantiation, centroid computation after feature extraction is linear in K and the number of pooled features; calibration scans a precomputed matrix in $O(G|\Lambda|)$ time.

3 EXPERIMENTS

We evaluate release utility, restorer and sampling sensitivity, and human agreement under CSG. DINO centroid is primary; $K = 8$ and $\alpha = 0.1$ are fixed before testing.

3.1 EXPERIMENTAL SETUP

The benchmark contains 1,000 masked instances from 695 Places365 sources at 256×256 resolution (Zhou et al., 2018). Development/calibration/test contain 201/402/397 instances from 139/278/278 disjoint sources; all 24 test instances without candidates remain in source averages. Mask areas span below 1%, 1–3%, 3–8%, and 8–15%. Restorers are SD-inpaint, a project-specific stochastic EAMamba, DDNM, and DPS (Rombach et al., 2022; Lin et al., 2025; Wang et al., 2023; Chung et al., 2023). SLIC uses 128 regions and $\rho = 0.05$ (Achanta et al., 2012); DINOv2-Small features and CLIP ViT-B/16 proxy labels follow Appendix B (Oquab et al., 2024; Radford et al., 2021). We separately assess SLIC target-granularity sensitivity for the primary SD-inpaint setting in Appendix D.5.

All scores share outputs, regions, proxy labels, source splits, and CRC; each has its own development grid and threshold. Table 1 compares RGB, LPIPS (Zhang et al., 2018), edge/line disagreement, regional PUQ/K-RCPS/Adame adaptations, and appearance/structure scores. Adaptations are not official reproductions. The 200 paired re-splits reuse one source collection; SD measures split variability. Fixed-test intervals use 5,000 source-bootstrap replicates conditional on the fitted threshold.

Risk averages L_g; region and candidate-area coverage average within sources then across sources. Conditional error averages sources with releases. We also report no-release rate, area under the

Table 1: **Regional score comparison on SD-natural.** Metrics averaged over sources; mean ± sample SD across 200 shared re-splits, $K = 8$, $\alpha = 0.10$; SD describes split variability. Bold/underline rank deployable-score means. PUQ/K-RCPS/Adame are regional adaptations. Table 7 reports additional fixed-test diagnostics.

Score	Proxy risk	Region coverage ↑	Area coverage ↑	Cond. error ↓
Conventional scores and random ranking				
RGB variance	0.097 ± 0.012	0.506 ± 0.040	0.421 ± 0.037	0.205 ± 0.011
Edge/line disagreement	0.097 ± 0.013	0.526 ± 0.039	0.567 ± 0.046	0.198 ± 0.012
LPIPS	0.097 ± 0.014	0.571 ± 0.049	0.530 ± 0.047	0.163 ± 0.012
Random ranking	0.096 ± 0.011	0.452 ± 0.026	0.451 ± 0.027	0.223 ± 0.012
Task-adapted uncertainty scores				
PUQ-derived	0.097 ± 0.012	0.486 ± 0.045	0.365 ± 0.042	0.209 ± 0.010
K-RCPS-derived	0.097 ± 0.013	0.495 ± 0.046	0.373 ± 0.043	0.207 ± 0.009
Adame-derived	0.097 ± 0.014	0.524 ± 0.052	0.513 ± 0.050	0.181 ± 0.013
Regional score instantiations within CSG				
DINO centroid (primary)	0.098 ± 0.013	0.653 ± 0.037	0.639 ± 0.040	0.162 ± 0.012
DINO spread	0.097 ± 0.013	0.651 ± 0.038	0.642 ± 0.039	0.162 ± 0.012
DINO pair	0.097 ± 0.013	0.667 ± 0.039	0.656 ± 0.041	$\mathbf{0.157} \pm 0.013$
Appearance fusion	0.098 ± 0.013	$\underline{0.670} \pm 0.036$	$\underline{0.661} \pm 0.039$	$\underline{0.158} \pm 0.013$
Appearance + structure	0.098 ± 0.012	$\mathbf{0.673} \pm 0.033$	$\mathbf{0.720} \pm 0.031$	0.159 ± 0.012
Reference-aware oracle	0.097 ± 0.010	0.846 ± 0.011	0.843 ± 0.011	0.136 ± 0.013

Table 2: **Within-restorer scores under shared SD-inpaint regions.** Mean ± SD over 200 recalibrated source re-splits ($K = 8$, $\alpha = 0.10$). Regions derive from the designated SD-inpaint output: controlled spatial comparisons, not native deployment performance. Native results and protocol: Appendix C.2.1. Split SDs differ from fixed-test EAMamba source-bootstrap CIs (5,000 draws).

Restorer	Score	Proxy risk	Region coverage	Area coverage	Conditional error	AURC
SD-inpaint	LPIPS	0.097 ± 0.014	0.571 ± 0.049	0.530 ± 0.047	0.163 ± 0.012	0.0798 ± 0.0049
	DINO spread	0.097 ± 0.013	0.651 ± 0.038	0.642 ± 0.039	0.162 ± 0.012	0.0689 ± 0.0043
	DINO centroid	0.098 ± 0.013	0.653 ± 0.037	0.639 ± 0.040	0.162 ± 0.012	0.0692 ± 0.0044
DDNM	LPIPS	0.099 ± 0.016	0.535 ± 0.034	0.512 ± 0.031	0.204 ± 0.022	0.0955 ± 0.0056
	DINO spread	0.099 ± 0.015	0.575 ± 0.034	0.544 ± 0.031	0.192 ± 0.019	0.0896 ± 0.0054
	DINO centroid	0.100 ± 0.016	0.563 ± 0.036	0.537 ± 0.035	0.193 ± 0.020	0.0909 ± 0.0054
DPS	LPIPS	0.098 ± 0.013	0.602 ± 0.027	0.553 ± 0.025	0.215 ± 0.022	0.0839 ± 0.0055
	DINO spread	0.098 ± 0.013	0.599 ± 0.026	0.570 ± 0.026	0.208 ± 0.016	0.0835 ± 0.0054
	DINO centroid	0.098 ± 0.012	0.599 ± 0.025	0.565 ± 0.025	0.214 ± 0.015	0.0836 ± 0.0054
EAMamba	LPIPS	0.097 ± 0.019	0.161 ± 0.025	0.136 ± 0.022	0.557 ± 0.030	0.3355 ± 0.0085
	DINO spread	0.097 ± 0.016	0.173 ± 0.021	0.124 ± 0.018	0.568 ± 0.034	0.3371 ± 0.0083
	DINO centroid	0.097 ± 0.015	0.170 ± 0.019	0.126 ± 0.017	0.603 ± 0.026	0.3390 ± 0.0084

risk–coverage curve (AURC), and area under the receiver operating characteristic curve (AUROC) for error discrimination. Bold/underlining mark the best/second displayed estimates within stated groups, including ties; arrows indicate direction. Ranks imply no significance, and stress summaries are descriptive.

3.2 Score Instantiations and Risk–Coverage Tradeoffs

Across 200 shared re-splits, the primary centroid instantiation attained coverage 0.653 ± 0.037, versus 0.486–0.571 for conventional and task-adapted scores; auxiliary spread was similar (0.651 ± 0.038; Table 1). Similar clean coverage need not imply equivalent response to designated-output changes

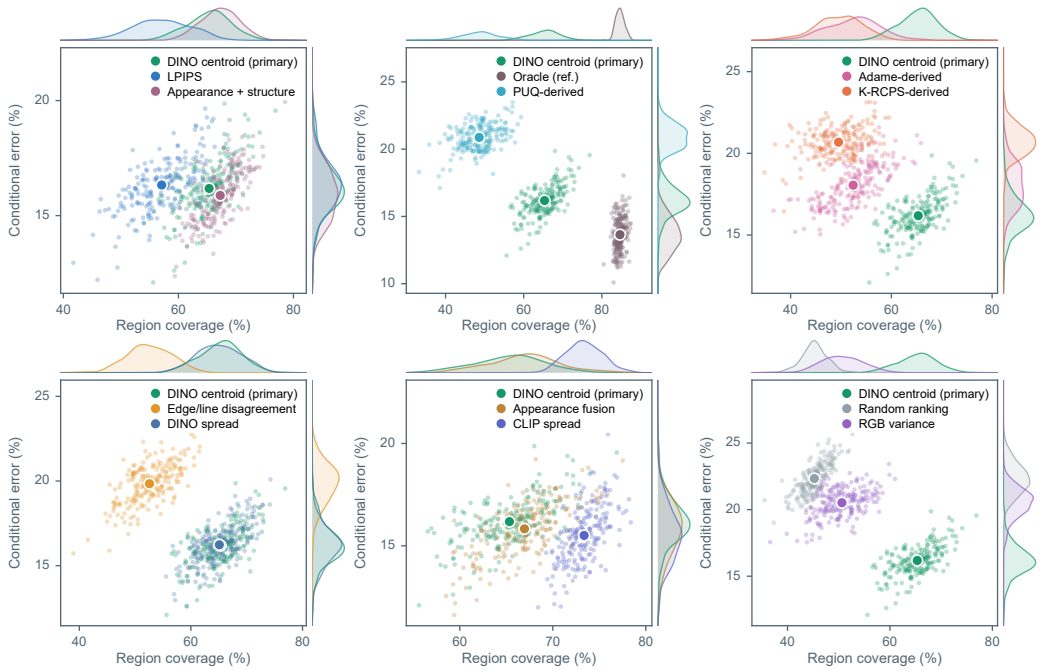

Figure 3: **Coverage and conditional proxy error across shared source re-splits.** At $K = 8$ and $\alpha = 0.10$, each panel pairs DINO centroid with two alternatives; the other 12 methods appear once. Small circles show all 200 re-splits per method; large circles mark means. Top and right curves show marginal kernel densities. Linear axes use panel-specific limits. Re-splits reuse one source collection. CLIP shares the proxy representation; the oracle uses intact references.

(Table 4). Adding structure to appearance fusion increased area coverage from 0.661 to 0.720, with conditional error 0.158 versus 0.159. Figure 3 shows these tradeoffs for 13 scores and diagnostics; Table 7 gives fixed-test points and AURC.

Pair mean attains coverage 0.667. CLIP scoring shares the proxy representation (Table 28(a)); human labels provide a separate target. Centroid has mean risk 0.098 but exceeds 0.1 in 43.5% of re-splits (Figure 2(a)). CRC controls marginal expected loss, not each split's risk.

3.3 Restorer, Sampling Stability, and Cost

Table 2 reports 200 re-split means and SDs under shared SD-inpaint regions. Centroid exceeds LPIPS mean coverage on SD-inpaint/DDNM, not DPS. On the original fixed test set, its coverage gains are 6.54/3.33/0.62 percentage points on SD-inpaint/DDNM/DPS with lower AURC (Appendix C.2).

EAMamba's fixed-test centroid remains in a low-coverage, high-error regime: native coverage/risk/conditional error/AURC are 0.158/0.082/0.521/0.331, versus shared 0.146/0.078/0.582/0.335. The conditional-error difference's paired interval includes zero, precluding an improvement claim. Released regions of at least 25 pixels retain source/instance-balanced proxy-error proportions 0.547/0.584 (shared/native), aggregated differently from conditional error. Similar median centroid scores for EAMamba and DPS accompany substantially different reference-proxy errors: sample agreement need not imply reference correctness. This limits the score instantiation without identifying a training, architectural, or implementation cause (Appendix G).

Partition sensitivity on the fixed split. Native-map rebuilds regions and recalibrates scores (Table 3). SD-inpaint is unchanged; DDNM changes little. DPS gains 3.13 coverage points [0.52, 5.74] with higher risk/error point estimates. DDNM/DPS centroid coverage ranks reverse; AURC rankings persist for all scores. Changed candidate counts and labels preclude geometry-only causal attribution (Appendix C.2.3).

Table 3: **DINO centroid: fixed-test partition comparison.** Original 278 test sources, not re-split means. Coverage: percent; Δ: Native−Shared, percentage points. Approximate simultaneous 95% intervals use 5,000 paired source-bootstrap draws across 45 contrasts. Candidate counts, geometry, and labels change together (Appendix C.2.3).

Restorer	Shared coverage	Native coverage	Δ coverage [95% interval]	Shared AURC	Native AURC
SD-inpaint	62.72	62.72	+0.00 [0.00, 0.00]	0.0681	0.0681
DDNM	58.10	58.20	+0.10 [-2.43, 2.63]	0.0924	0.0905
DPS	58.01	61.14	+3.13 [0.52, 5.74]	0.0795	0.0790
EAMamba	14.60	15.78	+1.18 [-1.26, 3.61]	0.3351	0.3313

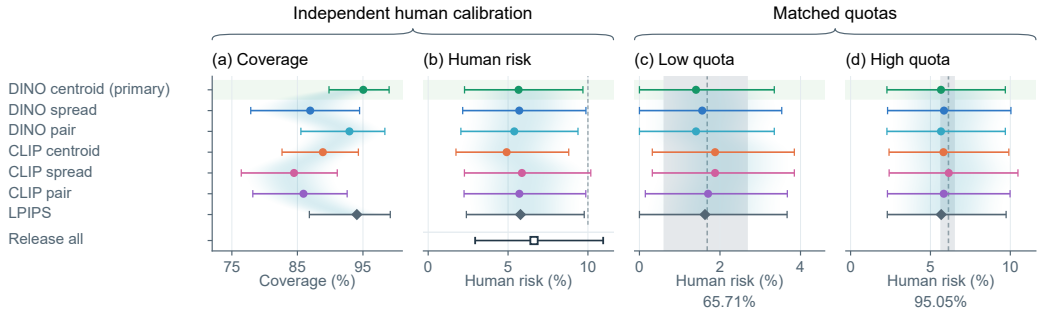

Figure 4: **Human calibration and ranking at two release quotas.** (a,b) Separately human-calibrated policies; release-all is uncalibrated. (c,d) Counts match proxy- and human-calibrated DINO centroid within each instance (65.7%/95.1% coverage). Test: 413 regions, 26 errors, 36 sources. Bars: 95% source-bootstrap intervals (5,000 draws), conditional on labels, split, thresholds, and quotas. Gray bands and dashed means: 1,000 random policies per quota. Risk scales differ; (b) marks the 10% budget.

Across $K = 2, 4, 8, 16$, coverage/AURC ranges are at most 0.0090/0.0015; sampling stability and generation/processing costs are reported separately in Appendices C.2 and H.4.

3.4 Human Validation and Distribution Shift

Human-audited utility and within-instance ranking. The human audit covers 955 regions of at least 25 pixels from 89 sources, with agreement 0.927 before adjudication. On 36 held-out sources, the original proxy threshold yielded human risk 0.014 [0.000, 0.033] and coverage 0.657 [0.532, 0.770] (Appendix E).

Under human calibration, centroid attained 95.1% coverage at risk 0.0566, versus LPIPS's 94.1%/0.0578 (Figure 4(a,b)). With matched within-instance quotas (panels (c,d)), centroid/LPIPS risks are 0.0140/0.0163 at 257 releases and 0.0566/0.0567 at 391 releases. The higher-quota paired interval includes zero; the lower quota lacks a paired interval. Random-policy distributions and full comparisons appear in Appendix H.2.

Release-all risk is 0.0663 [0.0294, 0.1096]. Fixed-denominator subset risks cannot exceed it empirically, so being below 0.1 cannot distinguish scores. An exploratory budget sweep gives 69.0%/70.6%/79.9% coverage for DINO centroid/pair/LPIPS at $\alpha = 0.05$. Budgets 0.01 and 0.025 trigger complete abstention with 35 calibration sources (Appendix I.2).

Response to changes in the designated output. Table 4(a) fixes auxiliaries, regions, and clean thresholds across nine release-only perturbations and four restorers. Spread and auxiliary-pair LPIPS remain unchanged; centroid changes 25.9% of decisions. Risk exceeds 0.1 in 0/36 centroid conditions, versus 27/36 for spread, 26/36 for LPIPS, and 2/36 for pair mean. Mask-wide perturbations test output sensitivity, not error localization.

After matched recalibration (panel (b)), centroid leaves fewer sources without releases than spread in 34/36 conditions (two ties) and LPIPS in all 36; pair mean retains higher coverage. Under color

Table 4: **Direct score comparison under release-only perturbations.** (a) Frozen clean thresholds; (b) matched recalibration; (c) matched color perturbations. Panels (a,b) average 36 conditions (four restorers $\times$ nine perturbations); (c) averages 12 color conditions. Risk is the source average of proxy burden; conditional error excludes sources without releases. Action changes use each clean threshold. Conditions share sources; exceedances are descriptive. Bold identifies the primary score.

(a) Frozen clean thresholds

Score	Mean risk	Max. risk	Risk > 0.1	Region coverage	Conditional error	Action changes
DINO centroid (primary)	**0.0441**	**0.0846**	**0/36**	**55.33%**	**22.85%**	**25.90%**
DINO spread	0.1642	0.4271	27/36	49.81%	37.00%	0.00%
DINO pair	0.0613	0.1021	2/36	31.27%	25.60%	19.37%
Auxiliary-pair LPIPS	0.1568	0.4134	26/36	46.04%	36.54%	0.00%

(b) Matched recalibration

Score	Mean risk	Region coverage	Conditional error	No-release sources	Risk > 0.1
DINO centroid (primary)	**0.0935**	**37.56%**	**31.86%**	**22.42%**	**10/36**
DINO spread	0.0955	36.52%	34.55%	33.60%	14/36
DINO pair	0.0948	38.86%	33.33%	21.95%	11/36
Auxiliary-pair LPIPS	0.0977	34.77%	35.61%	44.55%	15/36

(c) Color perturbations: matched recalibration

Score	Mean risk	Region coverage	No-release sources	Mean AUROC
DINO centroid (primary)	**0.0925**	**28.27%**	**31.86%**	**0.7589**
DINO spread	0.0922	21.72%	53.48%	0.6760
DINO pair	0.0924	28.23%	31.68%	0.7689
Auxiliary-pair LPIPS	0.0907	17.27%	64.81%	0.6195

perturbations (panel (c)), centroid coverage exceeds spread and LPIPS, while pair mean has similar coverage and higher AUROC.

Designated-to-auxiliary LPIPS also responds, changing 15.3% of decisions. Paired intervals for its human-risk differences from centroid include zero at both quotas (Appendix I.3).

Common-mode stress transforms designated and auxiliary outputs identically; 30/36 centroid conditions exceed 0.1 with clean thresholds (Appendix F). Neither arbitrary shifts nor human errors are covered by the guarantee.

3.5 Tiny-Region Audit and Mitigation

The audit sampled one region below 25 pixels per source, with 50 sources per geometry. Reviewers agreed on 77.5% ($\kappa = 0.587$). Excluding T153 leaves 199 records, including one unresolved label. The frozen policy releases 30, of which 19 are errors: 63.3% [45.2%, 80.0%], using geometry-stratified bootstrap resampling. This describes the selected stress sample (Figure 14; Appendix D).

On 320 test sources from a separate 800-source large-mask benchmark, 25-pixel guarding increases both area coverage and error burden; neighborhood pooling reduces both, but long thin risk remains 0.417 (Table 29(a)). Aggregate control does not ensure subgroup reliability.

4 Conclusion

CSG combines a designated output, separate scoring samples, a fixed error denominator, and existing CRC for regional delivery. Its guarantee concerns marginal expected proxy burden across sources. DINO centroid is one scoring choice; shared errors, limited feature support and subgroup imbalance constrain its utility.

AI USE STATEMENT

In accordance with the ICLR 2027 submission policy, we disclose that large language models (OpenAI ChatGPT) were used to assist with correcting grammar, editing LaTeX syntax, improving the visual presentation of tables and figures, visualizing results, and formatting the manuscript.

ETHICS STATEMENT

The authors report that annotators are specialists in cultural fields, received training, and participated with informed awareness. No formal ethics approval or exemption record was supplied for this revision. The human-label audit comprises 955 candidate regions from 89 source identities. It evaluates agreement between automatic proxy labels and human judgments, release utility under separate human-label calibration, and score ranking at matched within-instance release counts. A separate audit includes 200 tiny regions and 47 third-reviewer judgments. T153 is the assisted item; it is excluded from the primary analysis and included in a sensitivity analysis reported in Appendix H.3. Unresolved labels are retained.

REFERENCES

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012. doi: 10.1109/TPAMI.2012.120.

Eduardo Adame, Daniel Csillag, and Guilherme Tegoni Goedert. Image super-resolution with guarantees via conformalized generative models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Omer Belhasin, Yaniv Romano, Daniel Freedman, Ehud Rivlin, and Michael Elad. Principal uncertainty quantification with spatial correlation for image restoration problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3321–3333, 2024. doi: 10.1109/TPAMI.2023.3343031.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Noa Cohen, Hila Manor, Yuval Bahat, and Tomer Michaeli. From posterior sampling to meaningful diversity in image restoration. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ff2g30cZxj.

Santiago Cortes-Gomez, Carlos Patiño, Yewon Byun, Steven Wu, Eric Horvitz, and Bryan Wilder. Utility-directed conformal prediction: A decision-aware framework for actionable uncertainty quantification. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iOMnn1hSBO.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

António Farinhas, Chrysoula Zerva, Dennis Thomas Ulmer, and André F. T. Martins. Non-exchangeable conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=j511LaqEeP.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yuxiang Guo, Jiang Liu, Ze Wang, Hao Chen, Ximeng Sun, Yang Zhao, Jialian Wu, Xiaodong Yu, Zicheng Liu, and Emad Barsoum. ImageDoctor: Diagnosing text-to-image generation via grounded image reasoning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=04HwYGgp2w.

Kimia Hamidieh, Veronika Thost, Walter Gerych, Mikhail Yurochkin, and Marzyeh Ghassemi. Complementing self-consistency with cross-model disagreement for uncertainty quantification. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lOoRJo8xWy.

Eliahu Horwitz and Yedid Hoshen. Conffusion: Confidence intervals for diffusion models. *arXiv preprint arXiv:2211.09795*, 2022.

Hailey Joren, Charles Thomas Marx, and Berk Ustun. Classification with conceptual safeguards. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=t8cBsT9mcg.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

Klaus-Rudolf Kladny, Bernhard Schölkopf, and Michael Muehlebach. Conformal generative modeling with improved sample efficiency through sequential greedy filtering. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1i6lkavJ94.

Gilad Kutiel, Regev Cohen, Michael Elad, and Daniel Freedman. What's behind the mask: Estimating uncertainty in image-to-image problems. *arXiv preprint arXiv:2211.15211*, 2022. doi: 10.48550/arXiv.2211.15211.

Yu-Cheng Lin, Yu-Syuan Xu, Hao-Wei Chen, Hsien-Kai Kuo, and Chun-Yi Lee. EAMamba: Efficient all-around vision state space model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11708–11719, 2025.

Ségolène Martin, Anne Gagneux, Paul Hagemann, and Gabriele Steidl. PnP-Flow: Plug-and-play image restoration with flow matching. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=5AtHrq3B5R.

Jaewon Min, Jin Hyeon Kim, Paul Hyunbin Cho, Jaeeun Lee, Jihye Park, Minkyu Park, Sangpil Kim, Hyunhee Park, and Seungryong Kim. Text-aware image restoration with diffusion models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jt2c2H6auR.

Harikrishna Narasimhan, Aditya Krishna Menon, Wittawat Jitkrittum, and Sanjiv Kumar. Plugin estimators for selective classification with out-of-distribution detection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DASh78rJ7g.

Cuong Nguyen, Thanh-Toan Do, and Gustavo Carneiro. Probabilistic learning to defer: Handling missing expert annotations and controlling workload distribution. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=zl0HLZOJC9.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pzUhfQ74c5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

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*, pp. 10684–10695, 2022.

Jacopo Teneggi, Matthew Tivnan, Web Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33940–33960. PMLR, 2023.

Qingni Wang, Tiantian Geng, Zhiyuan Wang, Teng Wang, Bo Fu, and Feng Zheng. Sample then identify: A general framework for risk control and assessment in multimodal large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9WYMDgxDac.

Qingni Wang, Yue Fan, and Xin Eric Wang. SAFER: Risk-constrained sample-then-filter in large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kJmLmOvwLC.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=mRieQgMtNTQ.

Tongda Xu, Xiyan Cai, Xinjie Zhang, Xingtong Ge, Dailan He, Ming Sun, Jingjing Liu, Ya-Qin Zhang, Jian Li, and Yan Wang. Rethinking diffusion posterior sampling: From conditional score estimator to maximizing a posterior. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GcvLoqOoXL.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy T. Feng, Caifeng Zou, Yu Sun, Nikola Kovachki, Zachary E. Ross, Katherine L. Bouman, and Yisong Yue. Inversebench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=U3PBITXNG6.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.

Wenbin Zhou, Agni Orfanoudaki, and Shixiang Zhu. Conformalized decision risk assessment. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xRjOrcj08o.

Table 5: **Output objects and calibrated actions.** The calibration unit is an example unless stated otherwise. The Kutiel et al. (2022) column summarizes the method and calibration claims stated in the cited work. The CSG guarantee requires the assumptions in Section 2.4.

Field	Kutiel et al.	CSG
Output	Existing image prediction and a continuous spatial mask	Discrete candidate regions of a designated completion
Target	Distortion between masked prediction and reference	Released proxy errors divided by fixed candidate count
Unit	Image example under the calibration population	Source group containing one or more masked instances
Score inputs	Trained mask predictor using input and prediction	Frozen regional score; primary score compares designated and auxiliary features
Action	Attenuate pixels with a continuous mask	Release or abstain on missing candidate regions
Scope	Proposed conformal calibration of continuous masks; reference triples for training	Expected proxy burden under source exchangeability; no human or conditional error guarantee

A Method Details and Guarantee

This appendix specifies the protocol, implementation, extended results, annotation procedures, and observed limitations.

A.1 Positioning Relative to Neighboring Methods

Kutiel et al. calibrate continuous spatial masks for an existing prediction, including image completion (Kutiel et al., 2022). CSG uses discrete candidate regions and a released proxy error burden averaged within source identity. Table 5 contrasts these statistical objects. The CRC result applies existing theory. Adame et al. (2025) identify a gap and provide a counterexample for the guarantee of the related 2023 Conformal Prediction Masks method. Table 5 summarizes the 2022 formulation cited here. CSG's guarantee is derived independently from conformal risk control.

A.2 Score Identities and Diagnostic Variants

The primary instantiation uses the designated-to-centroid score from Equation (2). For compactness, write $d = \phi_{ir}(\hat{x}_i^{\mathrm{del}})$, $a_k = \phi_{ir}(\hat{x}_i^{\mathrm{aux},k})$, and $\bar{a} = K^{-1}\sum_{k=1}^{K} a_k$. The two disagreement scores are defined as

$$s_{ir}^{\mathrm{cent}} = \|d - \bar{a}\|_2^2, \qquad s_{ir}^{\mathrm{spr}} = \frac{1}{K}\sum_{k=1}^{K}\|a_k - \bar{a}\|_2^2,$$

where s_{ir}^{cent} measures the designated completion's deviation from the auxiliary center, and s_{ir}^{spr} measures dispersion within the auxiliary bank. The vector d represents the designated feature, a_k represents auxiliary feature k, and $\bar{a}$ is their mean. This centroid score is the primary instantiation within the CSG framework.

The mean squared designated-to-auxiliary distance decomposes into these two scores, as formulated in Equation (6):

$$\frac{1}{K}\sum_{k=1}^{K}\|d - a_k\|_2^2 = s_{ir}^{\mathrm{cent}} + s_{ir}^{\mathrm{spr}}, \tag{6}$$

where K is the auxiliary-bank size and the left-hand side averages the squared distances between the designated feature d and each a_k. The two terms on the right retain the centroid and dispersion definitions given above.

Proof. Expand $d - a_k = (d - \bar{a}) + (\bar{a} - a_k)$, average the squared norm over k, and use $\sum_k (a_k - \bar{a}) = 0$ to remove the cross term. □

A.3 Marginal proxy risk control across sources

We condition on every choice made with development data, including the regional scoring rule. Regions and repeated masks may be dependent within a source. The exchangeable calibration units are the source groups defined in Section 2.3.

Proof. For an instance with $R_i > 0$, every summand in Equation (3) is in $\{0, 1\}$, so $0 \le \ell_i(\lambda) \le 1$. The convention $\ell_i(\lambda) = 0$ when $R_i = 0$ preserves this bound. Because thresholded release sets are nested, $\lambda_1 \le \lambda_2$ implies

$$\mathbf{1}\{s_{ir} \le \lambda_1\} \le \mathbf{1}\{s_{ir} \le \lambda_2\}$$

for every candidate region. Hence $\ell_i(\lambda)$ is nondecreasing. Averaging all instances belonging to source identity g gives a source loss $L_g(\lambda) \in [0, 1]$ with the same monotonicity.

Reverse the parameter order by setting $\theta = -\lambda$ and $R_g(\theta) = L_g(-\theta)$. The resulting finite family is bounded and non-increasing. The largest feasible λ in Equation (4) is the smallest feasible θ after this reparameterization. Applying conformal risk control to the exchangeable source groups with uniform upper bound one yields

$$\mathbb{E}[R_{G+1}(\hat{\theta})] \le \alpha,$$

where R_{G+1} denotes the reparameterized loss of a future source and $\hat{\theta} = -\hat{\lambda}$ denotes the calibrated threshold in reversed order. This bound is equivalent to $\mathbb{E}[L_{G+1}(\hat{\lambda})] \le \alpha$.

The grid is finite, and ties are resolved by taking the last feasible grid point. The sentinel $-\infty$ corresponds to an empty release set. If the finite-sample correction makes every grid point infeasible, CSG returns $\hat{\lambda} = \perp$ and abstains everywhere; the resulting loss is zero. The analysis permits dependence among regions and repeated instances from the same source. Exchangeability is required across source groups under the frozen within-source protocol. □

This direct application of CRC establishes marginal control of proxy risk across sources for any frozen regional scoring rule that induces the bounded nested loss family defined in Section 2.3. Grouping by source accommodates dependence among regions and repeated instances within each source.

A.4 Calibration and Deployment Algorithms

Algorithms 1 and 2 make explicit the separation between choices during development, calibration across sources, and deployment. With a cached region map, the primary centroid instantiation uses $K + 1$ restorer calls and $K + 1$ score-encoder passes; proxy encodings are required only for reference-available development and calibration pairs. Without the shared map, non-SD restorers additionally require the designated SD output and its regional partition.

B Implementation and Reproducibility

This section specifies the implementation used to evaluate the guarantee, including the score, calibration object, sampling cost, and uncertainty estimates.

B.1 Frozen Implementation Contract

Table 6 records the choices that determine the score, proxy target, and calibration object. These protocol choices remain fixed throughout calibration and evaluation.

B.2 Analytic Cost

For the primary centroid instantiation, an uncached input requires one restorer call for the a priori designated completion and K calls for the auxiliary bank. It then performs $K+1$ score-encoder passes, which may be batched when memory permits. Reference- available development and calibration additionally require proxy encodings of the designated completion and intact reference. After feature extraction, centroid construction and regional pooling are linear in K and in the number of pooled region-feature values. Scanning a precomputed source-loss matrix costs $O(G|\Lambda|)$.

Algorithm 1 Source-group calibration for CSG

Require: Frozen restorer and sampling rules, regionizer, regional scoring rule S, score representation, intact references, source IDs and instance membership, proxy encoder, proxy threshold τ, threshold grid Λ, risk budget α, calibration source groups

for each calibration instance i **do**
  Load or generate $\hat{x}_i^{\mathrm{del}}$ and $\{\hat{x}_i^{\mathrm{aux},k}\}_{k=1}^{K}$
  Load shared SD regions or generate the designated SD output and region map
  if $R_i = 0$ **then**
    Set $\ell_i(\lambda) = 0$ for all $\lambda \in \Lambda$
  else
    Compute frozen regional scores s_{ir}
    Compute proxy labels from the designated completion and reference
    Evaluate $\ell_i(\lambda)$ over the frozen threshold grid
  end if
end for
for each calibration source identity g **do**
  Average all associated instance losses to obtain $L_g(\lambda)$
end for
Compute $\widehat{B}_G(\lambda)$ on Λ
if at least one grid point is feasible **then**
  return the largest feasible threshold $\hat{\lambda}$
else
  return $\perp$ (all ABSTAIN)
end if

Algorithm 2 CSG release for one restoration instance

Require: Masked input Y_i, frozen protocol, calibrated threshold $\hat{\lambda}$

Generate or load one a priori designated completion and K auxiliary completions
Load shared SD regions or generate the designated SD output and region map
if $\hat{\lambda} = \perp$ or $R_i = 0$ **then**
  Apply ABSTAIN to the entire missing support
else
  Compute the inputs required by the frozen regional scoring rule
  for each $r \in \mathcal{R}_i$ **do**
    Compute s_{ir}
    if $s_{ir} \le \hat{\lambda}$ **then**
      RELEASE the designated-completion pixels in r
    else
      Apply the visible ABSTAIN rendering in r
    end if
  end for
  Apply ABSTAIN to uncovered missing pixels
end if

B.3 Resampling and Uncertainty Estimates

Main operating-point intervals use 5,000 source-bootstrap replicates. Calibration reliability and paired score comparisons use 200 source-grouped re-splits. The human audit uses source-clustered bootstrap intervals, and the stress audit uses 5,000 source-bootstrap replicates per restorer and condition. The Monte Carlo interval around a re-split mean and the percentile range across individual splits estimate different quantities and are labeled separately.

B.4 Calibration Scope

The a priori designated-output rule, auxiliary bank, regional scoring rule and representation, proxy contract, regionizer, feature processing, and deployment population jointly define one calibrated protocol. A change to any component establishes a new protocol and receives a corresponding

Table 6: **Primary SD-natural implementation contract.**

Object	Frozen specification
Statistical unit	Source identity. The 201/402/397 instances correspond to 139/278/278 development/calibration/test sources. All instances from one source stay together.
A priori designated object	The per-instance manifest fixes one seed/index before scoring under the internal field `delivery`; the same object is reused across all K, which precludes output-dependent selection.
Auxiliary bank	Disjoint manifest roles `aux_01`,...,`aux_K`; primary $K = 8$. The K-sweep uses nested prefixes up to 16, and the stability audit uses three independent replicate banks.
Regionizer	SLIC with 128 targets, overlap threshold $\rho = 0.05$, connected components of the strict mask intersection, and pooling restricted to the missing support.
Score encoder	DINOv2-Small (`facebook/dinov2-small`; fallback bundle `dinov2_vits14`). A 256×256 completion is resized to 252×252 and ImageNet-normalized. The classification (CLS) token is removed, leaving an 18×18 patch grid.
Score pooling	The dense DINOv2 feature map is bilinearly interpolated to the region-map resolution and mean-pooled over each masked-intersection region. Interpolation assigns pooled features to every retained region, including regions between native patch centers. Appendix D analyzes the resulting scale behavior.
Proxy encoder	CLIP ViT-B/16 (`openai/clip-vit-base-patch16`), resized to 224×224 with CLIP normalization. The CLS token is removed, the 14×14 patch grid is bilinearly interpolated, and regional features are mean-pooled.
Proxy target	Cosine distance between designated-completion and intact-reference CLIP embeddings. The frozen threshold is the 0.75 development quantile, $\tau = 0.13966859463872094$.
Calibration grid	1,001 development-score quantiles plus $-\infty$ and $+\infty$ sentinels; ties select the largest feasible grid point. An empty feasible set returns all ABSTAIN.
Zero-region rule	Instance loss and coverage are both zero before within-source averaging. No-release fraction and area coverage complete the utility accounting.
Rendering	Observed pixels are preserved; RELEASE regions display the a priori designated completion from the frozen manifest, while uncovered or rejected missing pixels receive a visible ABSTAIN pattern.
Cost	One designated plus K auxiliary restorer calls, followed by $K + 1$ score-encoder passes. Reference-available calibration additionally requires proxy features for the designated output and reference.

calibration threshold. The large-mask and common-mode experiments quantify this transfer and the effect of matched recalibration.

The remaining sections evaluate this fixed protocol. We first report the complete quantitative extensions and then examine geometry, expert labels, and systematic perturbations.

B.5 Complete Score and Preprocessing Definitions

Regions and fixed preprocessing. SLIC is computed on the designated SD-inpaint completion in RGB $[0, 1]$, with 128 requested segments, compactness 10, and start label 0. Each superpixel is intersected with the missing mask; the retained overlap fraction is at least 0.05, and four-connected components define candidates. The resulting cached regions are reused across restorers and scores. Release-only perturbations retain this same region map. Thus neither the intact reference nor a perturbed output is used to rebuild candidates. Dense feature maps are interpolated to the image grid before averaging over candidate pixels (OpenCV bilinear interpolation for DINO/CLIP maps). No pooled-vector unit normalization is applied to squared DINO/CLIP distances.

DINO, CLIP, and LPIPS. For either encoder, centroid, spread, and pair scores use the definitions in Section 2.2, with the specified encoder substituted for ϕ. The LPIPS score in Tables 1 and 4 is the mean spatial LPIPS-Alex v0.1 map over all $\binom{8}{2} = 28$ auxiliary pairs, averaged within each candidate. Inputs are rescaled to $[-1, 1]$, the model is in evaluation mode, and its spatial maps are bilinearly

resized to the image grid. The additional designated-to-auxiliary score uses

$$s_{ir}^{\text{LPIPS-del}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{A_{ir}}\sum_{p\in r} D_{\text{LPIPS}}(\hat{x}_i^{\text{del}}, \hat{x}_i^{\text{aux},k})(p), \qquad A_{ir} = |r|.$$

Both use the same regional pooling and LPIPS representation; only access to the designated output differs.

Pixelwise uncertainty adaptations. The following maps use all eight auxiliary completions. Regional aggregation, fixed during development, takes the maximum over candidate pixels for PUQ and K-RCPS, and the mean for Adame before CSG calibration. These rules were selected from mean, 90th percentile, and maximum by development source-macro AURC, choosing the first rule in that order within 0.001 of the minimum. For K-RCPS-derived scoring, convert each pixel to the mean of its three RGB channels divided by 255; the map is its empirical 0.95 quantile minus its 0.05 quantile across the eight samples, using linear interpolation. For PUQ-derived scoring, center the eight RGB vectors in $[0,1]^3$ at each pixel and obtain principal directions v_j by SVD. If w_j is the 0.95-minus-0.05 quantile width of their projections onto v_j, the map is $\max_{c\in\{R,G,B\}}\sum_j |v_{cj}|w_j$. All available RGB directions are used; no auxiliary subsampling is applied. For Adame-derived scoring, transform RGB to Lab, normalize channels by $(100, 254, 254)$ and add $(0, 0.5, 0.5)$. Compute the population variance across auxiliaries in each channel, average the three variances, and apply a Gaussian filter with $\sigma = 32$ pixels and reflect boundary handling. These are regional adaptations, not reproductions of the original release objects or guarantees.

Appearance and structure fusion. Let F_q be the empirical CDF of score q fitted on development data, assigning equal total weight to each development source that contains candidate regions and equal weight to its regions. Appearance fusion is $\max\{F_{\text{cent}}(s^{\text{cent}}), F_{\text{spr}}(s^{\text{spr}})\}$. Appearance plus structure additionally takes the maximum with $F_{\text{edge}}(s^{\text{edge}})$. For the edge score, form grayscale as the RGB mean and use 3×3 Sobel gradients to obtain magnitude m and angle θ. The pixel feature is $(\bar{m}, \bar{m}\cos 2\theta, \bar{m}\sin 2\theta)$, where $\bar{m} = m/\max(\max_p m(p), 10^{-6})$ within each image. After regional pooling, s^{edge} is auxiliary spread in this three-dimensional feature space. RGB variance averages channelwise population variances across the eight auxiliary images, then across region pixels. Random ranking uses a frozen NumPy generator; the oracle uses the proxy distance computed from the reference and is not deployable.

Stochastic EAMamba. The evaluated EAMamba is a project-specific backbone in a residual-diffusion inpainting pipeline. Its integer seed resets Python, NumPy, PyTorch, and CUDA generators before each completion, controlling diffusion initial noise and any stochastic trajectory draws. Image resizing, mask feathering, line extraction, prompt, checkpoint, and 50-step sampling schedule remain fixed across seeds. The complete historical prompt string and checkpoint identity require recovery from generation records. Line conditioning is extracted once per instance and reused for every seed. Observed pixels are projected back exactly after generation. This protocol does not inject dropout into a deterministic EAMamba restorer.

Development sources in repeated splits. The original 139 development sources remain excluded from calibration and test in every one of the 200 re-splits. Only the original calibration/test pool is repartitioned by source identity, with the same assignment for all scores. Score normalization, proxy threshold, and score grids remain fitted to the original development set; each re-split fits only its calibration threshold.

C Extended Quantitative Results

This section extends the primary operating point through score diagnostics, calibration sweeps, cross-restorer evaluation, and regional case analysis. The quantitative comparisons are followed by examples of regional release decisions.

Table 7: **Fixed-test score diagnostics on SD-natural.** Fixed test: 278 sources, 397 instances; $K = 8$, $\alpha = 0.10$. All values are fractions averaged over sources. Bold/underline rank all deployable rows together; the reference-aware oracle is excluded. PUQ-, K-RCPS-, and Adame-derived scores are regional adaptations, not official reproductions. CLIP spread shares the proxy-label representation.

Score	Proxy risk	Region coverage ↑	Area coverage ↑	Cond. error ↓	AURC ↓
RGB variance	0.100	0.511	0.427	0.209	0.089
Edge/line disagreement	0.105	0.554	0.594	0.200	0.087
LPIPS	0.099	0.562	0.516	0.172	0.080
Random ranking	0.082	0.409	0.405	0.214	0.090
PUQ-derived	0.104	0.506	0.385	0.211	0.093
K-RCPS-derived	0.106	0.525	0.402	0.200	0.093
Adame-derived	0.104	0.537	0.529	0.193	0.087
DINO centroid (primary)	0.093	0.627	0.610	0.151	0.068
DINO spread	0.103	<u>0.663</u>	0.658	0.159	0.068
Appearance fusion	0.094	0.651	0.640	0.156	<u>0.066</u>
Appearance + structure	0.088	0.642	<u>0.685</u>	<u>0.140</u>	0.067
CLIP spread	0.084	**0.702**	**0.713**	**0.136**	**0.053**
Reference-aware oracle	0.081	0.831	0.826	0.117	0.019

C.1 Additional SD-natural Results

C.1.1 Score diagnostics

Table 7 complements the repeated-split results in Table 1 with fixed-test operating points and AURC on 278 sources. These values use a single calibrated threshold per score. CLIP spread is included here as a proxy-aligned diagnostic, while the oracle uses the intact reference. Their proxy-based results should be interpreted separately from the human-error comparisons.

The primary centroid score evaluates the completion eligible for release. Auxiliary spread omits that completion, whereas the archived appearance and appearance-plus-structure fusions combine related score components. Their fixed-test results quantify the utility tradeoff between these constructions.

C.1.2 Proxy-risk-budget sweep

Table 8(a) reports the fixed test point and the 200 source-grouped re-split distribution. The proxy-risk split range reports the empirical 2.5–97.5 percentile interval across individual splits.

Increasing α increases coverage and proxy risk. From 0.01 to 0.20, DINO centroid region coverage on the fixed test split rises from 0.128 to 0.931. Mean proxy risk across the source re-splits tracks the selected budget.

C.1.3 Calibration-source count

For each row in Table 9, source identities are sampled without replacement and all associated instances are retained. The full $G = 278$ endpoint is deterministic.

Increasing the calibration set consistently expands coverage at both risk budgets. At $\alpha = 0.1$, region coverage rises from 0.421 with 20 sources to 0.627 with all 278 sources, while proxy risk approaches the selected budget.

C.1.4 Zero-region and no-release accounting

The fixed test population contains 397 restoration instances from 278 source identities. Twenty-four instances contain no retained candidate region and contribute zero loss and zero coverage before source aggregation. At the primary $\alpha = 0.1$ operating point, 256 of 278 sources release at least one region; 22 sources have undefined error conditional on release and are excluded only from

Table 8: **Sensitivity to the proxy-risk budget.** (a) SD-natural diagnostics: fixed-test points and 200 shared re-split summaries; Exceed. is the observed fraction above the budget. (b) Four-restorer fixed-test results, with risk/region coverage in each cell. All budgets use the primary paired-score split assignments. (a) Coverage is ranked between scores at each shared budget; (b) operating points at different budgets and restorers are descriptive.

(a) SD-natural fixed-test and re-split diagnostics

Score	α	Fixed risk	Fixed region coverage ↑	Re-split mean risk	Re-split risk range	Exceed.
DINO centroid (primary)	0.01	0.0111	0.1284	0.0067	[0.0019, 0.0128]	0.170
DINO centroid (primary)	0.025	0.0258	0.2176	0.0215	[0.0104, 0.0347]	0.255
DINO centroid (primary)	0.05	0.0464	0.3380	0.0469	[0.0308, 0.0647]	0.355
DINO centroid (primary)	0.075	0.0717	0.5049	0.0721	[0.0522, 0.0945]	0.390
DINO centroid (primary)	0.10	0.0927	0.6272	0.0976	[0.0739, 0.1215]	0.435
DINO centroid (primary)	0.15	0.1328	**0.8145**	0.1476	[0.1183, 0.1767]	0.420
DINO centroid (primary)	0.20	0.1860	**0.9311**	0.1962	[0.1649, 0.2212]	0.410
CLIP diagnostic	0.01	0.0091	**0.1819**	0.0067	[0.0017, 0.0143]	0.120
CLIP diagnostic	0.025	0.0201	**0.3398**	0.0223	[0.0105, 0.0341]	0.315
CLIP diagnostic	0.05	0.0424	**0.5370**	0.0472	[0.0306, 0.0648]	0.385
CLIP diagnostic	0.075	0.0617	**0.6324**	0.0723	[0.0524, 0.0917]	0.410
CLIP diagnostic	0.10	0.0840	**0.7025**	0.0971	[0.0773, 0.1171]	0.380
CLIP diagnostic	0.15	0.1311	0.8039	0.1472	[0.1230, 0.1745]	0.420
CLIP diagnostic	0.20	0.1821	0.9170	0.1964	[0.1641, 0.2212]	0.445

(b) Fixed-test risk / region coverage across restorers

α	SD-inpaint	EAMamba	DDNM	DPS
0.01	0.011/0.128	0.011/0.035	0.010/0.165	0.004/0.165
0.025	0.026/0.218	0.025/0.060	0.034/0.302	0.015/0.274
0.05	0.046/0.338	0.046/0.097	0.066/0.426	0.044/0.431
0.075	0.072/0.505	0.061/0.122	0.087/0.508	0.068/0.522
0.10	0.093/0.627	0.078/0.146	0.108/0.581	0.089/0.580
0.15	0.133/0.815	0.131/0.215	0.154/0.700	0.143/0.711
0.20	0.186/0.931	0.177/0.272	0.210/0.813	0.189/0.798

Table 9: **Calibration-size sensitivity.** Values are means over 200 source subsamples, except the full 278-source endpoint. Coverage extrema are ranked across calibration sizes within each budget; larger calibration sets use more data.

	$\alpha = 0.05$			$\alpha = 0.10$		
G	Proxy risk	Region coverage ↑	Area coverage ↑	Proxy risk	Region coverage ↑	Area coverage ↑
20	0.0082	0.0968	0.0845	0.0587	0.4207	0.4009
40	0.0312	0.2497	0.2275	0.0752	0.5230	0.5057
80	0.0390	0.2952	0.2714	0.0852	0.5834	0.5672
120	0.0419	0.3126	0.2883	0.0882	0.6009	0.5846
160	0.0441	0.3247	0.3001	0.0903	0.6127	0.5964
220	0.0461	0.3359	0.3121	0.0920	0.6210	0.6048
278	0.0464	**0.3380**	**0.3140**	0.0927	**0.6272**	**0.6105**

that descriptive ratio. The controlled macro-average of proxy risk across sources and the coverage calculation retain all 278 sources.

Area coverage, the no-release count, and the explicit zero-region convention provide complete utility accounting under the theorem's fixed denominator.

C.1.5 BASELINE IMPLEMENTATION BOUNDARY

The simple baselines are computed directly on the frozen candidate regions. RGB variance is the regional mean of per-pixel auxiliary RGB variance. LPIPS uses the mean spatial LPIPS-Alex map over all 28 auxiliary pairs. CLIP and edge diagnostics use auxiliary spread in their respective feature spaces. Random ranking assigns a fixed pseudorandom scalar per region, and the oracle uses the designated-to-reference proxy distance.

PUQ, K-RCPS, and Adame are evaluated as frozen region-level task adaptations under the same CRC loss. This shared evaluation isolates the regional ranking signal while holding the deployment policy fixed.

C.2 CROSS-RESTORER AND SAMPLING RESULTS

C.2.1 REPEATED-SPLIT UNCERTAINTY

Table 2 replaces the original fixed-test point estimates with means and sample SDs over the same 200 source-grouped re-splits as Table 1. The 556 non-development sources are repartitioned into 278 calibration and 278 test sources, retaining all instances of a source together. The 139 development sources, development grids, restoration samples, and proxy cutoff remain fixed. Each score/restorer/map protocol is recalibrated at $\alpha = 0.1$ in each repetition. SD describes split variability, not a population confidence interval; the repetitions reuse sources. Table 11 reports the native-map repetitions. The fixed-test comparisons below and the EAMamba diagnosis instead use the original 397 test instances from 278 sources and source-bootstrap intervals conditional on fitted thresholds. Their point estimates must not be compared with re-split means as if partition geometry were the only difference.

Having established the SD-natural operating behavior, we apply the same release protocol to EAMamba, DDNM, and DPS. The analyses separate restorer-specific coverage from the shared calibration behavior. DINO centroid released nothing for 53.2% of EAMamba test sources versus 7.9% for SD-inpaint; DDNM exceeded the nominal risk budget on this fixed split.

C.2.2 COMMON PROTOCOL AND STOCHASTICITY CHECKS

In the shared-map evaluation, all four restorers use the same 1,000 instances, source-disjoint split, designated-plus-eight output roles, region maps, DINOv2 score representation, CLIP proxy, and frozen proxy threshold. Each restorer is calibrated separately. The EAMamba implementation was qualified on 100 sources with 16 outputs per source. Repeated calls with the same seed were identical, observed pixels were preserved, and all 100 sources exhibited masked-output diversity above the frozen threshold. DDNM and DPS each supplied the complete 9,000-image designated-plus-eight cache.

C.2.3 SHARED VERSUS NATIVE SPATIAL PARTITIONS

Shared-map uses candidate regions from the designated SD-inpaint completion for all restorers. Native-map rebuilds regions from each restorer's own designated output. Both retain the 1,000 instances, the frozen 139/278/278 development/calibration/test source split, designated outputs, eight auxiliary completions, feature extractors, and proxy cutoff. SLIC uses 128 targets, compactness 10, and overlap threshold 0.05. Each restorer, score, and map protocol has its own development grid and calibrated threshold at $\alpha = 0.1$. No restoration outputs are resampled. SD-inpaint partitions are identical; shared-map metrics reproduce the original results.

The following tables report all 24 operating points with pointwise 95% percentile intervals and all 60 Native$-$Shared metric differences with approximate simultaneous 95% intervals. All intervals use 5,000 paired resamples of the 278 test sources and are conditional on fitted thresholds, development grids, outputs, and proxy labels. Regions are not paired across protocols. AURC integrates source-average proxy error burden against region coverage on the development grid; each bootstrap draw reconstructs this curve before integration.

The simultaneous family comprises three non-SD restorers, three scores, and five metrics: 45 contrasts. The critical value, 3.2170, is the 95th percentile of the maximum absolute standardized, centered bootstrap deviation across this family. Each interval is the observed difference plus or minus this

Table 10: **Auxiliary-count sensitivity and sampling stability.** (a) Primary-bank fixed-test results at $\alpha = 0.1$ for 16 restorer/K settings. (b) Three-bank stability on the 100-source subset: Spearman and flip are mean ± sample SD; other columns are means. Coverage SD uses the primary-bank threshold. The designated completion is fixed across K. Each input needs $K + 1$ restorer calls and score-encoder passes: 3, 5, 9, and 17 for $K = 2, 4, 8, 16$. Ranks compare the four K values within each restorer and panel; only means are highlighted when SDs are reported.

(a) Primary-bank operating points

Restorer	K	Proxy risk	Region coverage ↑	Area coverage ↑	Cond. error ↓	AURC ↓	No release ↓
SD-inpaint	2	0.0907	0.6211	0.6056	0.1502	**0.0681**	0.0827
SD-inpaint	4	0.0916	**0.6287**	0.6100	0.1491	**0.0681**	**0.0719**
SD-inpaint	8	0.0927	0.6272	**0.6105**	0.1510	**0.0681**	0.0791
SD-inpaint	16	0.0896	0.6207	0.6043	**0.1475**	0.0684	0.0935
EAMamba	2	0.0793	0.1452	0.1059	0.5518	0.3366	0.5612
EAMamba	4	0.0773	0.1445	**0.1091**	**0.5499**	0.3357	0.5432
EAMamba	8	0.0775	**0.1460**	0.1059	0.5815	**0.3351**	**0.5324**
EAMamba	16	0.0747	0.1397	0.0996	0.5529	0.3361	0.5504
DDNM	2	0.1135	0.5801	0.5577	0.2011	0.0927	0.1655
DDNM	4	0.1060	0.5809	0.5517	**0.1836**	**0.0924**	**0.1511**
DDNM	8	0.1085	0.5810	0.5558	0.1889	**0.0924**	0.1583
DDNM	16	0.1142	**0.5861**	**0.5614**	0.1978	0.0929	0.1583
DPS	2	0.0922	**0.5877**	**0.5546**	0.2090	0.0800	**0.0683**
DPS	4	0.0840	0.5787	0.5507	0.2036	**0.0795**	0.0791
DPS	8	0.0891	0.5801	0.5491	0.2033	**0.0795**	0.0899
DPS	16	0.0883	0.5842	0.5529	**0.1997**	0.0796	0.0899

(b) Auxiliary-bank stability

Restorer	K	Spearman ↑	Score CV ↓	Action agreement ↑	Release Jaccard ↑	Flip rate ↓	Coverage SD ↓
SD-inpaint	2	0.849 ± 0.009	0.400	0.860	0.773	0.140 ± 0.003	0.0066
SD-inpaint	4	0.906 ± 0.004	0.321	0.898	0.808	0.102 ± 0.007	0.0139
SD-inpaint	8	0.945 ± 0.001	0.250	0.909	0.836	0.091 ± 0.007	0.0145
SD-inpaint	16	**0.970** ± 0.001	**0.174**	**0.939**	**0.883**	**0.061** ± 0.005	**0.0055**
EAMamba	2	0.866 ± 0.013	0.313	0.924	0.804	0.076 ± 0.016	0.0074
EAMamba	4	0.895 ± 0.010	0.265	0.940	0.799	0.060 ± 0.004	**0.0047**
EAMamba	8	0.945 ± 0.011	0.196	0.962	0.878	0.038 ± 0.006	0.0084
EAMamba	16	**0.968** ± 0.006	**0.146**	**0.974**	**0.902**	**0.026** ± 0.006	0.0082
DDNM	2	0.941 ± 0.004	0.520	0.893	0.758	0.107 ± 0.004	0.0031
DDNM	4	0.951 ± 0.004	0.455	0.893	0.766	0.107 ± 0.004	0.0144
DDNM	8	0.971 ± 0.004	0.346	0.931	0.823	0.069 ± 0.013	**0.0021**
DDNM	16	**0.984** ± 0.004	**0.261**	**0.948**	**0.882**	**0.052** ± 0.006	0.0046
DPS	2	0.918 ± 0.003	0.368	0.873	0.744	0.127 ± 0.008	**0.0055**
DPS	4	0.950 ± 0.005	0.296	0.888	0.773	0.112 ± 0.002	0.0101
DPS	8	0.969 ± 0.003	0.227	0.927	0.834	0.073 ± 0.007	0.0115
DPS	16	**0.984** ± 0.001	**0.169**	**0.949**	**0.874**	**0.051** ± 0.003	0.0068

critical value times its bootstrap standard deviation. SD-inpaint differences are identically zero. These intervals do not include calibration-sample or annotation uncertainty and are not confidence intervals for the CRC guarantee.

For DINO centroid, DPS coverage increases by 3.13 percentage points, with simultaneous interval [0.52, 5.74]. Its risk rises from 0.0891 to 0.0968 and conditional error from 0.2033 to 0.2133; their

Table 11: **Within-restorer score comparisons under restorer-native spatial partitions.** Mean ± sample SD across the same 200 source-grouped calibration/test re-splits as Table 1. Each split uses 278 calibration and 278 test sources; all 139 development sources and development grids remain fixed. Thresholds are recalibrated separately for each score, restorer, and map protocol at $K = 8$ and $\alpha = 0.10$. SD measures split-to-split variability. Test instance counts vary by split. The EAMamba diagnostic analysis instead uses the original fixed test set (397 instances from 278 sources), frozen thresholds, and 5,000 source-bootstrap replicates for 95% confidence intervals; those intervals are not the SDs reported here. AURC integrates source-macro fixed-denominator proxy risk against region coverage.

Restorer	Score	Proxy risk	Region coverage	Area coverage	Conditional error	AURC
SD-inpaint	LPIPS	0.097 ± 0.014	0.571 ± 0.049	0.530 ± 0.047	0.163 ± 0.012	0.0798 ± 0.0049
	DINO spread	0.097 ± 0.013	0.651 ± 0.038	0.642 ± 0.039	0.162 ± 0.012	0.0689 ± 0.0043
	DINO centroid	0.098 ± 0.013	0.653 ± 0.037	0.639 ± 0.040	0.162 ± 0.012	0.0692 ± 0.0044
DDNM	LPIPS	0.098 ± 0.016	0.536 ± 0.036	0.510 ± 0.032	0.198 ± 0.021	0.0939 ± 0.0054
	DINO spread	0.099 ± 0.015	0.562 ± 0.035	0.528 ± 0.032	0.204 ± 0.017	0.0885 ± 0.0052
	DINO centroid	0.099 ± 0.016	0.563 ± 0.036	0.534 ± 0.035	0.187 ± 0.019	0.0896 ± 0.0053
DPS	LPIPS	0.098 ± 0.012	0.607 ± 0.026	0.559 ± 0.025	0.217 ± 0.021	0.0827 ± 0.0051
	DINO spread	0.098 ± 0.012	0.608 ± 0.025	0.576 ± 0.025	0.214 ± 0.018	0.0819 ± 0.0050
	DINO centroid	0.097 ± 0.011	0.606 ± 0.020	0.577 ± 0.020	0.216 ± 0.016	0.0823 ± 0.0051
EAMamba	LPIPS	0.096 ± 0.018	0.165 ± 0.024	0.141 ± 0.023	0.554 ± 0.033	0.3297 ± 0.0086
	DINO spread	0.097 ± 0.015	0.176 ± 0.022	0.129 ± 0.019	0.536 ± 0.033	0.3317 ± 0.0082
	DINO centroid	0.097 ± 0.014	0.175 ± 0.019	0.136 ± 0.017	0.566 ± 0.027	0.3337 ± 0.0082

Table 12: **Shared and native operating points: SD-inpaint.** Each estimate is followed by a pointwise 95% interval from 5,000 source-bootstrap resamples. All values are fractions. Protocol-difference simultaneous intervals are reported separately.

Score	Map	Coverage	Proxy risk	Cond. error	AURC	Area cov.
DINO centroid	Shared	0.6272 [0.5886, 0.6664]	0.0927 [0.0759, 0.1109]	0.1510 [0.1254, 0.1784]	0.0681 [0.0570, 0.0807]	0.6105 [0.5690, 0.6525]
DINO centroid	Native	0.6272 [0.5886, 0.6664]	0.0927 [0.0759, 0.1109]	0.1510 [0.1254, 0.1784]	0.0681 [0.0570, 0.0807]	0.6105 [0.5690, 0.6525]
DINO spread	Shared	0.6628 [0.6240, 0.7014]	0.1027 [0.0854, 0.1223]	0.1586 [0.1325, 0.1860]	0.0679 [0.0569, 0.0803]	0.6577 [0.6174, 0.6977]
DINO spread	Native	0.6628 [0.6240, 0.7014]	0.1027 [0.0854, 0.1223]	0.1586 [0.1325, 0.1860]	0.0679 [0.0569, 0.0803]	0.6577 [0.6174, 0.6977]
Auxiliary-pair LPIPS	Shared	0.5617 [0.5151, 0.6096]	0.0987 [0.0797, 0.1208]	0.1719 [0.1436, 0.2036]	0.0801 [0.0665, 0.0953]	0.5162 [0.4677, 0.5663]
Auxiliary-pair LPIPS	Native	0.5617 [0.5151, 0.6096]	0.0987 [0.0797, 0.1208]	0.1719 [0.1436, 0.2036]	0.0801 [0.0665, 0.0953]	0.5162 [0.4677, 0.5663]

difference intervals include zero. EAMamba coverage remains low (0.1460 versus 0.1578), and conditional error remains high (0.5815 versus 0.5210). Thus, changing the partition does not remove its poor operating performance. DDNM has small point-estimate changes; intervals containing zero do not establish equivalence.

Coverage ordering changes: DDNM and DPS exchange positions for centroid and spread. For each of the three scores, AURC point estimates retain the order SD-inpaint, DPS, DDNM, EAMamba from lowest to highest under both protocols. This ordering does not establish pairwise statistical separation or unchanged risk–coverage curves. Native-map also changes candidate counts, geometry, and regional proxy labels. The ablation therefore measures sensitivity to the complete partitioning protocol and does not isolate boundary alignment as the sole cause.

C.2.4 Proxy-risk-budget sweep across restorers

Each cell in Table 8(b) reports "proxy risk/region coverage" on the fixed test split. The sweep evaluates how the same calibrated release rule adapts to each restorer's score distribution.

Table 13: **Shared and native operating points: DDNM.** Each estimate is followed by a pointwise 95% interval from 5,000 source-bootstrap resamples. All values are fractions. Protocol-difference simultaneous intervals are reported separately.

Score	Map	Coverage	Proxy risk	Cond. error	AURC	Area cov.
DINO centroid	Shared	0.5810 [0.5343, 0.6291]	0.1085 [0.0861, 0.1327]	0.1889 [0.1560, 0.2236]	0.0924 [0.0768, 0.1091]	0.5558 [0.5076, 0.6061]
DINO centroid	Native	0.5820 [0.5359, 0.6295]	0.1079 [0.0865, 0.1307]	0.1879 [0.1541, 0.2211]	0.0905 [0.0752, 0.1065]	0.5524 [0.5029, 0.6030]
DINO spread	Shared	0.5999 [0.5549, 0.6456]	0.1133 [0.0912, 0.1373]	0.2013 [0.1683, 0.2357]	0.0912 [0.0760, 0.1074]	0.5664 [0.5188, 0.6158]
DINO spread	Native	0.5869 [0.5404, 0.6348]	0.1117 [0.0906, 0.1342]	0.1959 [0.1625, 0.2302]	0.0898 [0.0747, 0.1054]	0.5521 [0.5018, 0.6021]
Auxiliary-pair LPIPS	Shared	0.5477 [0.4979, 0.5985]	0.1073 [0.0849, 0.1318]	0.2122 [0.1739, 0.2511]	0.0970 [0.0809, 0.1141]	0.5255 [0.4746, 0.5785]
Auxiliary-pair LPIPS	Native	0.5527 [0.5037, 0.6028]	0.1080 [0.0860, 0.1326]	0.2084 [0.1702, 0.2476]	0.0948 [0.0791, 0.1112]	0.5235 [0.4730, 0.5752]

Table 14: **Shared and native operating points: DPS.** Each estimate is followed by a pointwise 95% interval from 5,000 source-bootstrap resamples. All values are fractions. Protocol-difference simultaneous intervals are reported separately.

Score	Map	Coverage	Proxy risk	Cond. error	AURC	Area cov.
DINO centroid	Shared	0.5801 [0.5373, 0.6248]	0.0891 [0.0722, 0.1077]	0.2033 [0.1706, 0.2372]	0.0795 [0.0658, 0.0946]	0.5491 [0.5039, 0.5952]
DINO centroid	Native	0.6114 [0.5683, 0.6543]	0.0968 [0.0802, 0.1148]	0.2133 [0.1798, 0.2470]	0.0790 [0.0653, 0.0937]	0.5862 [0.5409, 0.6314]
DINO spread	Shared	0.5769 [0.5340, 0.6211]	0.0901 [0.0726, 0.1098]	0.1884 [0.1572, 0.2220]	0.0801 [0.0660, 0.0954]	0.5496 [0.5030, 0.5974]
DINO spread	Native	0.6016 [0.5584, 0.6456]	0.0968 [0.0801, 0.1162]	0.2123 [0.1790, 0.2475]	0.0795 [0.0658, 0.0943]	0.5725 [0.5269, 0.6209]
Auxiliary-pair LPIPS	Shared	0.5738 [0.5265, 0.6222]	0.0864 [0.0683, 0.1064]	0.1909 [0.1546, 0.2268]	0.0804 [0.0658, 0.0960]	0.5327 [0.4830, 0.5846]
Auxiliary-pair LPIPS	Native	0.6004 [0.5553, 0.6480]	0.0945 [0.0766, 0.1144]	0.2121 [0.1766, 0.2491]	0.0804 [0.0662, 0.0955]	0.5558 [0.5073, 0.6073]

Table 15: **Shared and native operating points: EAMamba.** Each estimate is followed by a pointwise 95% interval from 5,000 source-bootstrap resamples. All values are fractions. Protocol-difference simultaneous intervals are reported separately.

Score	Map	Coverage	Proxy risk	Cond. error	AURC	Area cov.
DINO centroid	Shared	0.1460 [0.1196, 0.1740]	0.0775 [0.0605, 0.0965]	0.5815 [0.5137, 0.6483]	0.3351 [0.3122, 0.3577]	0.1059 [0.0825, 0.1303]
DINO centroid	Native	0.1578 [0.1306, 0.1891]	0.0819 [0.0629, 0.1032]	0.5210 [0.4514, 0.5897]	0.3313 [0.3081, 0.3550]	0.1200 [0.0949, 0.1480]
DINO spread	Shared	0.1527 [0.1246, 0.1829]	0.0834 [0.0650, 0.1033]	0.5574 [0.4932, 0.6201]	0.3347 [0.3121, 0.3567]	0.1095 [0.0851, 0.1359]
DINO spread	Native	0.1615 [0.1330, 0.1935]	0.0810 [0.0627, 0.1014]	0.4718 [0.4046, 0.5391]	0.3290 [0.3065, 0.3521]	0.1169 [0.0915, 0.1455]
Auxiliary-pair LPIPS	Shared	0.1581 [0.1258, 0.1942]	0.0954 [0.0711, 0.1229]	0.5628 [0.4936, 0.6319]	0.3355 [0.3123, 0.3584]	0.1339 [0.1029, 0.1682]
Auxiliary-pair LPIPS	Native	0.1634 [0.1303, 0.2014]	0.0929 [0.0690, 0.1200]	0.5363 [0.4643, 0.6107]	0.3298 [0.3061, 0.3534]	0.1415 [0.1096, 0.1779]

All four restorers exhibit increasing coverage as the risk budget grows. At $\alpha = 0.1$, proxy risk ranges from 0.078 to 0.109, while coverage reflects the output distribution of each base restorer.

C.2.5 Full auxiliary-count sweep

The designated completion, region map, and proxy labels remain fixed as K changes. Each row in Table 10(a) receives a separate source-group CRC threshold.

Table 16: **Native−Shared differences: SD-inpaint.** Differences and approximate simultaneous 95% intervals from 5,000 paired source-bootstrap resamples, adjusted across the 45 non-SD contrasts. All values are fractions; SD-inpaint differences are identically zero.

Score	Coverage	Proxy risk	Cond. error	AURC	Area cov.
DINO centroid	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]
DINO spread	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]
Auxiliary-pair LPIPS	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]

Table 17: **Native−Shared differences: DDNM.** Differences and approximate simultaneous 95% intervals from 5,000 paired source-bootstrap resamples, adjusted across the 45 non-SD contrasts. All values are fractions; SD-inpaint differences are identically zero.

Score	Coverage	Proxy risk	Cond. error	AURC	Area cov.
DINO centroid	+0.0010 [-0.0243, 0.0263]	-0.0005 [-0.0149, 0.0139]	-0.0010 [-0.0397, 0.0377]	-0.0018 [-0.0111, 0.0074]	-0.0034 [-0.0278, 0.0210]
DINO spread	-0.0129 [-0.0391, 0.0133]	-0.0015 [-0.0173, 0.0142]	-0.0054 [-0.0501, 0.0393]	-0.0014 [-0.0105, 0.0076]	-0.0143 [-0.0380, 0.0095]
Auxiliary-pair LPIPS	+0.0050 [-0.0191, 0.0291]	+0.0007 [-0.0132, 0.0147]	-0.0038 [-0.0508, 0.0432]	-0.0022 [-0.0112, 0.0068]	-0.0019 [-0.0244, 0.0206]

Table 18: **Native−Shared differences: DPS.** Differences and approximate simultaneous 95% intervals from 5,000 paired source-bootstrap resamples, adjusted across the 45 non-SD contrasts. All values are fractions; SD-inpaint differences are identically zero.

Score	Coverage	Proxy risk	Cond. error	AURC	Area cov.
DINO centroid	+0.0313 [0.0052, 0.0574]	+0.0077 [-0.0068, 0.0221]	+0.0100 [-0.0284, 0.0484]	-0.0004 [-0.0083, 0.0074]	+0.0371 [0.0123, 0.0619]
DINO spread	+0.0247 [0.0013, 0.0481]	+0.0067 [-0.0082, 0.0217]	+0.0239 [-0.0170, 0.0648]	-0.0006 [-0.0085, 0.0074]	+0.0229 [0.0009, 0.0448]
Auxiliary-pair LPIPS	+0.0266 [0.0058, 0.0474]	+0.0081 [-0.0060, 0.0221]	+0.0212 [-0.0228, 0.0652]	-0.0000 [-0.0078, 0.0077]	+0.0232 [0.0044, 0.0420]

Table 19: **Native−Shared differences: EAMamba.** Differences and approximate simultaneous 95% intervals from 5,000 paired source-bootstrap resamples, adjusted across the 45 non-SD contrasts. All values are fractions; SD-inpaint differences are identically zero.

Score	Coverage	Proxy risk	Cond. error	AURC	Area cov.
DINO centroid	+0.0118 [-0.0126, 0.0361]	+0.0044 [-0.0095, 0.0182]	-0.0606 [-0.1615, 0.0404]	-0.0038 [-0.0208, 0.0132]	+0.0141 [-0.0089, 0.0371]
DINO spread	+0.0088 [-0.0165, 0.0340]	-0.0024 [-0.0136, 0.0088]	-0.0857 [-0.1898, 0.0185]	-0.0057 [-0.0213, 0.0099]	+0.0074 [-0.0146, 0.0295]
Auxiliary-pair LPIPS	+0.0053 [-0.0175, 0.0282]	-0.0025 [-0.0146, 0.0096]	-0.0265 [-0.0901, 0.0371]	-0.0057 [-0.0195, 0.0081]	+0.0076 [-0.0136, 0.0289]

Across all four restorers, the principal risk and coverage metrics remain stable from $K = 2$ to $K = 16$. For SD-inpaint, region coverage stays between 0.621 and 0.629, and the corresponding ranges are similarly compact for the other restorers. We use $K = 8$ because utility changes little with larger banks, while each instance then requires nine completions; $K = 16$ requires seventeen.

C.2.6 SEED-BANK STABILITY

Three disjoint replicate auxiliary banks are evaluated on a 100-source subset, with thresholds frozen from the primary bank. Table 10(b) summarizes changes in rank, release actions, and coverage when the auxiliary bank is replaced.

At $K = 8$, Spearman correlation ranges from 0.945 to 0.971 and action agreement ranges from 0.909 to 0.962. The across-bank SD of region coverage remains at or below 0.0145 across all restorer/K settings.

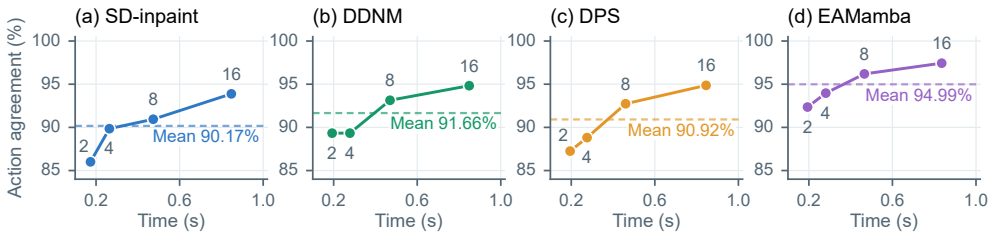

Figure 5: **Auxiliary-bank stability versus DINO centroid processing time.** (a–d) SD-inpaint, DDNM, DPS, and EAMamba. Point labels give $K \in \{2, 4, 8, 16\}$. Action agreement averages three disjoint auxiliary banks on 100 sources, with thresholds frozen from the primary bank. Dashed lines show each restorer's unweighted mean across the four K values. Time averages 100 cached-input measurements per setting after 20 warm-ups, excluding image generation, disk input, and decoding. Larger banks improve agreement while increasing processing cost.

Figure 5 combines these stability estimates with the separately measured mean DINO centroid processing time. Labels beside points give $K \in \{2, 4, 8, 16\}$; larger banks improve action agreement but require more processing and restorer calls.

The mean-agreement lines summarize the evaluated settings. They do not provide confidence bounds. Appendix H.4 specifies the separately measured processing components.

The current paired SD-natural encoder–score re-splits are reported in Table 28(a). Cross-restorer operating points and bank stability above use their specified fixed test populations.

C.3 Qualitative Decision Cases

The quantitative results above compare release policies at the population level. Figure 6 makes the corresponding regional actions visible for eight test-set examples and four restorers. Each masked input is held fixed across the four model columns so that the designated completion and the resulting regional action can be inspected under the same visual context. Their population frequencies are quantified by the aggregate analyses above.

D Mask-Geometry and Region Diagnostics

This section reports a comparison on one region record per source and diagnostics on the complete regional policy. These evaluations use different region populations, targets, and threshold protocols.

D.1 Evaluation populations, targets, and thresholds

The illustrative comparison uses constructed scenario records, 50 per geometry, with one stored region per nominal source. These records are not empirical annotations and do not support the main experimental claims. Each geometry contains 13, 13, 12, and 12 records in the four mask-area bins. The input is this fixed list of region records; no SLIC candidate set is reconstructed or sampled for this comparison. Coverage describes the listed regions. Area coverage averages the recorded released-region area divided by mask area.

The scenario supplies 0/1/U labels by construction. Label risk is the number of released records labeled 1 divided by all records in the evaluated group. The U record is always withheld and remains in the denominator; it is not recoded as label 0. This target is distinct from the CLIP proxy defined in Section 2.3 and does not use the proxy threshold in Table 6. No CLIP proxy labels are inferred from these labels.

Two fixed score thresholds, 35 and 20.924512, define nested release sets. No calibration-source fit is supplied for either threshold in this comparison. The comparison reports their numeric values. The second value matches the threshold in the separate tiny-region audit, but this equality does not transfer its calibration provenance or guarantee.

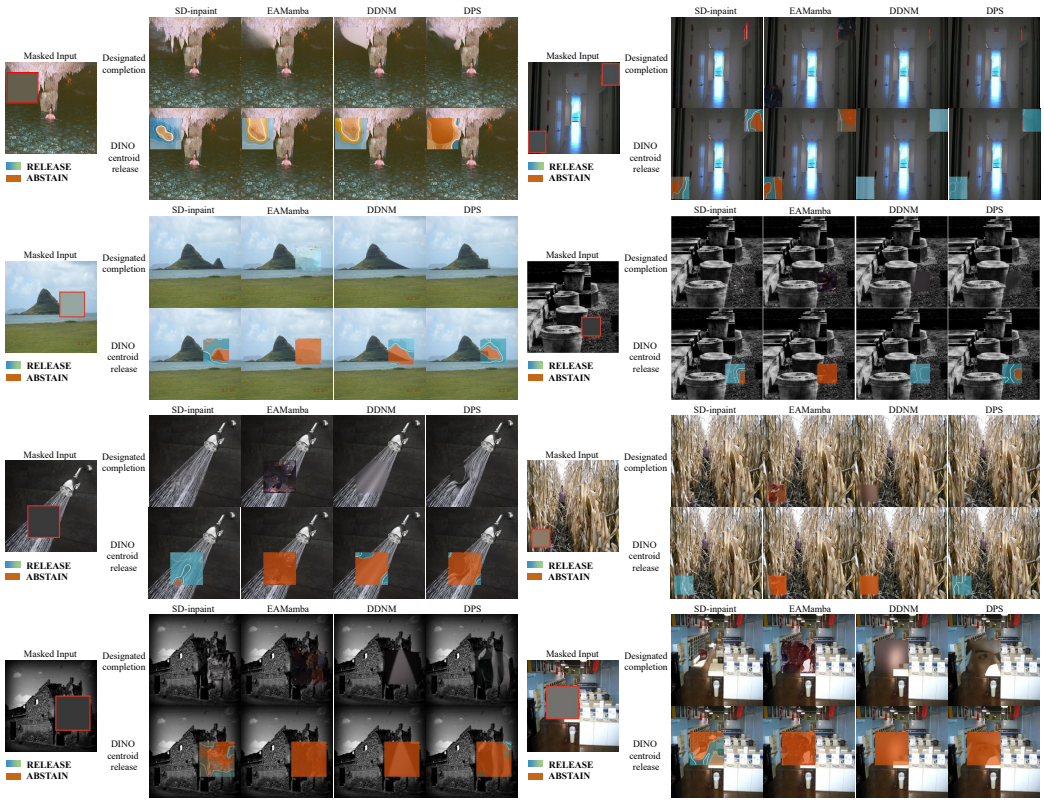

Figure 6: **Qualitative regional decisions across restorers.** Eight test-set examples compare SD-inpaint, EAMamba, DDNM, and DPS under the shared regional delivery protocol. Each block shows the masked input, the designated completion fixed *a priori*, and the corresponding DINO centroid release decision. Blue-green regions receive RELEASE and orange regions receive ABSTAIN. Contours visualize the frozen region scores.

The benchmark for the complete policy contains 800 sources partitioned into 160 development, 320 calibration, and 320 test sources. Its proxy risk and coverage results use the 320-source test split and the CLIP proxy. The tiny-region human audit in Section 3.5 is a separate 200-item evaluation; its annotations and release counts are not pooled with the comparison of region records. The evaluations use separate record tables and serve different purposes. Their source correspondence is not documented.

D.2 Operating points and region configurations

Table 20(a) reports the operating points for region records. Panel (b) evaluates complete candidate sets on the 320-source test split; each region configuration has its own calibrated thresholds. The SAM-based region configuration uses the original Segment Anything Model (SAM) (Kirillov et al., 2023). Intersection configurations pool within missing support, while the full-q control includes observed pixels. The two panels do not form a matched comparison.

In the comparison of region records, tightening the threshold from 35 to 20.924512 changes label risk from 0.330 to 0.145, region coverage from 0.635 to 0.315, and area coverage from 0.467 to 0.231. The SLIC-128 results for the complete policy in panel (b) refer to a different evaluation population.

D.3 Geometry and mask-area subgroups

Table 21(a) stratifies the 200 single-region records. Panel (b) summarizes all candidate regions in the separate 800-source benchmark. A median of 538.5 regions per long thin source in (b) describes the complete candidate sets, not the one-record inputs of (a). The panels use different populations and targets.

Table 20: **Operating points for region records and the complete policy.** (a) Constructed scenario records, 200 nominal sources; label risk uses the supplied labels. (b) Complete candidate sets on 320 test sources; proxy risk uses the CLIP target. Columns in (b) transfer the small-mask threshold or use the threshold estimated on 320 large-mask calibration sources. Panels have different populations and targets.

(a) Region records: 200 sources

	Threshold 35			Threshold 20.924512		
Setting	Label risk	Region coverage	Area coverage	Label risk	Region coverage	Area coverage
One region/source	0.330	0.635	0.467	0.145	0.315	0.231

(b) Complete regional policy: 320 test sources

	Small-to-large			Large recalibrated		
Configuration	Proxy risk	Region coverage	Area coverage	Proxy risk	Region coverage	Area coverage
SLIC-64, $\rho = 0.05$	0.241	0.329	0.309	0.104	0.157	0.182
SLIC-128, $\rho = 0.05$	0.238	0.315	0.292	0.104	0.151	0.173
SLIC-256, $\rho = 0.05$	0.224	0.290	0.268	0.105	0.148	0.162
SLIC-128, $\rho = 0.01$	0.227	0.299	0.277	0.104	0.152	0.174
SLIC-128, $\rho = 0.25$	0.274	0.366	0.347	0.105	0.153	0.175
Matched grid, $\rho = 0.05$	0.220	0.288	0.272	0.107	0.150	0.149
Matched SAM, $\rho = 0.05$	0.256	0.321	0.363	0.109	0.152	0.191
Full-q leakage control	0.571	0.830	0.809	0.103	0.182	0.188

Table 21: **Strata of region records and fragmentation of complete sets.** (a) 200 constructed scenario records; label risk uses constructed labels and no-release refers to threshold 20.924512. (b) Complete candidate sets from the separate 800-source benchmark. The panels use different region populations.

(a) Single-region records: 200 sources

	Threshold 35			Threshold 20.924512			
Stratum	Label risk	Region coverage	Area coverage	Label risk	Region coverage	Area coverage	No-release
Geometry strata							
Compact irregular	0.160	0.560	0.419	0.060	0.220	0.168	0.780
Long thin	0.580	0.900	0.673	0.380	0.620	0.453	0.380
Multiple components	0.240	0.500	0.352	0.080	0.160	0.113	0.840
Boundary intersecting	0.340	0.580	0.425	0.060	0.260	0.191	0.740
Mask-area strata							
15-30%	0.327	0.654	0.470	0.135	0.346	0.249	0.654
30-50%	0.308	0.654	0.481	0.173	0.346	0.248	0.654
50-70%	0.396	0.708	0.522	0.167	0.312	0.227	0.688
$> 70\%$	0.292	0.521	0.394	0.104	0.250	0.198	0.750

(b) Complete candidate sets: 800-source benchmark

Geometry	Regions per source	Region area	Token count	Zero-token fraction	Aspect ratio	Score
Compact irregular	59.0	495.0	2.0	0.207	1.39	490.6
Long thin	538.5	10.0	0.0	0.901	2.67	14.9
Multiple components	89.5	158.0	1.0	0.462	1.40	368.2
Boundary intersecting	54.5	556.0	3.0	0.272	1.33	555.5

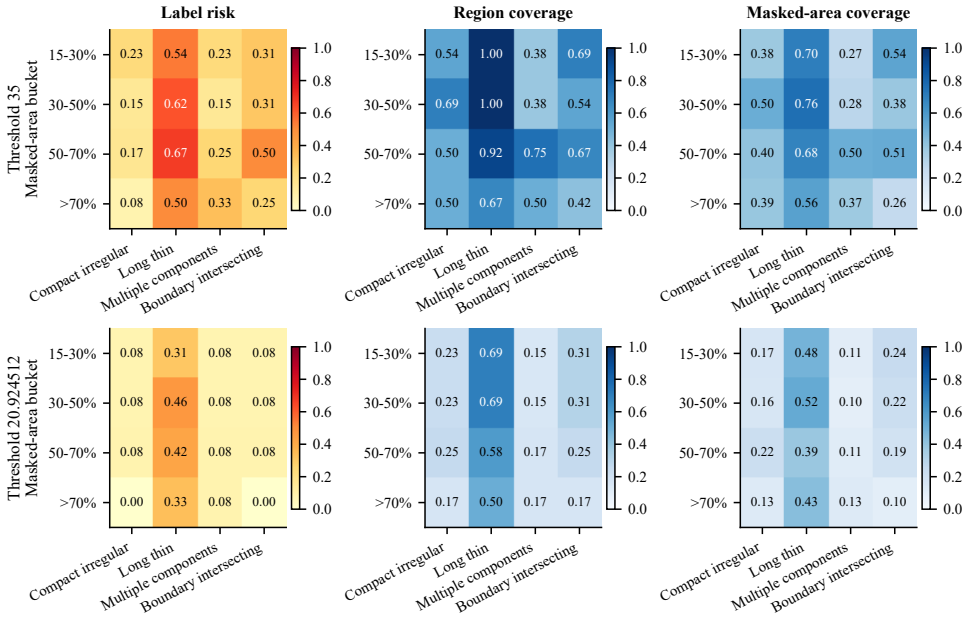

Figure 7: **Mask-area and geometry stratification of constructed scenario records.** Rows use fixed thresholds 35 and 20.924512. Columns show label risk, record coverage, and area coverage. Each geometry has 50 sources distributed 13/13/12/12 across area bins; subgroup aggregates use these counts.

Label risk and coverage decrease across all four area strata between the two operating points. All geometries retain releases: region coverage is 0.220, 0.620, 0.160, and 0.260, respectively. Long thin records have the highest label risk (0.380) and the largest single contribution to released area, approximately 49.0%. The other three groups have label risks of 0.060, 0.080, and 0.060.

D.4 Region-scale diagnosis

The complete-set diagnostic analyzes 1,978,876 region rows under the original score and operating point. Long thin masks create hundreds of small components. Table 21(b) relates this geometry to region area and DINOv2 token support.

Long thin masks have a median of 538.5 regions per source and a median region area of 10 pixels. Their zero-token fraction of 0.901 highlights the reliance of small regions on interpolated features for regional scoring. Compact irregular and boundary intersecting regions retain larger areas and greater token support, producing the complementary score regimes shown in Figure 8.

D.5 SLIC target-granularity sensitivity

The regional delivery protocol depends on the candidate partition. We therefore evaluate sensitivity to the SLIC target segment count in the primary SD-inpaint setting. We use 1,000 instances and consider target counts $N \in \{64, 128, 256\}$. Each configuration is recalibrated independently using the same source-level protocol as the main experiment, with $K = 8$, $\alpha = 0.10$, SLIC compactness 10.0, and overlap cutoff 0.05. Table 22 reports mean and standard deviation over 200 source-grouped calibration/test re-splits for DINO centroid and DINO spread.

For DINO centroid, mean proxy risk remains close to the target budget across the three target counts, at 0.099 ± 0.015, 0.098 ± 0.013, and 0.098 ± 0.012 for $N = 64$, 128, and 256, respectively. The operational quantities vary with the partition: coverage decreases from 0.664 ± 0.041 to 0.614 ± 0.034, conditional error increases from 0.151 ± 0.013 to 0.170 ± 0.012, and AURC increases from 0.0618 ± 0.0046 to 0.0802 ± 0.0042 as the target count increases. DINO spread follows the same qualitative pattern.

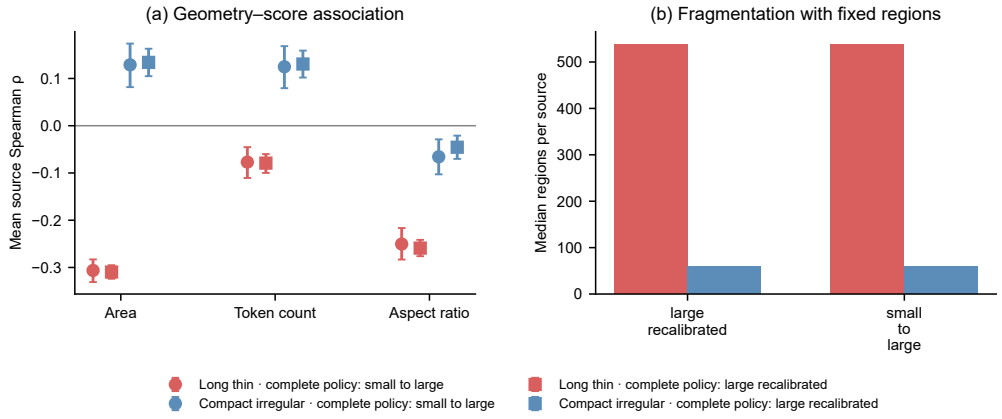

Figure 8: **Region-scale diagnostics.** All quantities refer to the separate benchmark for the complete policy. The figure relates disagreement to pixel area, DINOv2 token support, aspect ratio, and fragmentation for the frozen region configurations, characterizing associations between geometry and score behavior.

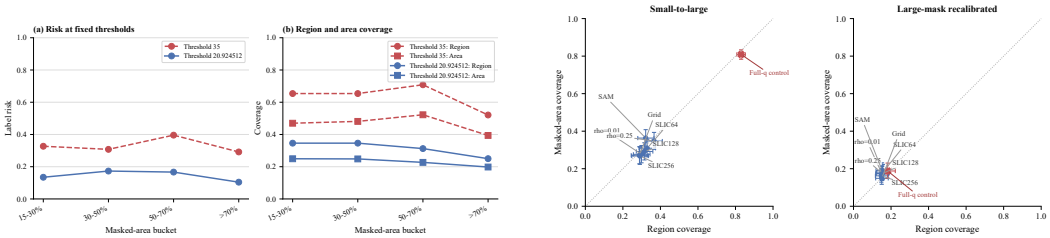

Figure 9: **Constructed scenario records and empirical complete-policy summaries.** Left: the 200-record comparison, stratified by mask area and evaluated at two fixed thresholds; no uncertainty intervals are supplied. Right: results for the complete policy on 320 test sources with the original intervals. The panels correspond to Table 20(a) and (b), respectively, and are not matched. SAM denotes the matched SAM configuration in Table 20(b).

Table 22: Sensitivity to SLIC target granularity on SD-inpaint. Each target count is separately calibrated using 200 source-grouped calibration/test re-splits. Values are mean ± standard deviation. Changing N changes candidate-region geometry, the candidate population, and represented damaged area.

Target N	Score	Coverage	Proxy risk	Cond. error	AURC
64	DINO centroid	0.664 ± 0.041	0.099 ± 0.015	0.151 ± 0.013	0.0618 ± 0.0046
128	DINO centroid	0.653 ± 0.037	0.098 ± 0.013	0.162 ± 0.012	0.0692 ± 0.0044
256	DINO centroid	0.614 ± 0.034	0.098 ± 0.012	0.170 ± 0.012	0.0802 ± 0.0042
64	DINO spread	0.657 ± 0.037	0.098 ± 0.015	0.151 ± 0.015	0.0617 ± 0.0046
128	DINO spread	0.651 ± 0.038	0.097 ± 0.013	0.162 ± 0.012	0.0689 ± 0.0043
256	DINO spread	0.624 ± 0.034	0.097 ± 0.012	0.168 ± 0.012	0.0794 ± 0.0041

These results show that source-level calibration retains the intended proxy-risk scale across the tested granularities, whereas release utility depends on the candidate-region definition. The comparison does not identify a universally preferred target count: changing N changes region geometry, the candidate-region population, and the damaged area represented by those candidates. In particular, the lower AURC at $N = 64$ should be interpreted within its own candidate-region protocol rather than as direct evidence that coarser segmentation dominates finer segmentation.

E HUMAN VALIDATION

The human study evaluates whether the proxy-calibrated policy also maintains low risk under expert judgments. Table 23(d) provides the primary validation result. On the held-out expert-labeled split, proxy calibration retains substantial region and area coverage while keeping the human risk averaged over sources low.

E.1 ITEM CONSTRUCTION AND INTERFACE

The formal audit contains 955 candidate regions from 89 source identities; regions smaller than 25 pixels are excluded. Each item displays the masked observation, designated completion, and intact reference, with matched full-image and crop views and an identical region outline. The model name, auxiliary samples, primary centroid score, automatic proxy label, calibration threshold, routing action, other responses, and source identity are hidden.

Annotators answer whether the designated restoration substantially disagrees with the intact reference within the marked region. Responses are 0 (agreement within tolerance), 1 (substantial disagreement), or U (unable to judge). Positive items may be tagged as semantic/object, geometry/topology, boundary continuity, color/illumination, texture/material, or artifact/seam errors, with severity 1–4 and confidence 1–5.

Table 23(a) records the source and item composition of the development, calibration, and held-out annotation splits.

E.2 ASSIGNMENT AND ADJUDICATION

Three primary annotators each receive 700 presentations. Every unique item is rated independently by two annotators; 10.0% repeated presentations assess intra-rater consistency and are separated by at least 100 positions. A 30-item pilot confirms the rubric before the formal run. All 70 primary rating differences are adjudicated by a fourth annotator. The final labels include 114 expert-confirmed positive regions.

Table 23(c) reports agreement for the primary ratings and the repeated presentations.

Across all items, raw pairwise agreement is 0.927, Fleiss' κ is 0.568, and nominal Krippendorff's α is 0.569. The raw and chance-corrected measures together support consistent application of the annotation rubric.

E.3 EXPERT ALIGNMENT OF THE PROXY

Table 23(b) documents the alignment between automatic proxy labels and final expert labels over all audited regions.

The overall proxy AUROC is 0.743 with a 95% source bootstrap interval [0.691, 0.789], and AUPRC is 0.203 [0.132, 0.277]. The AUPRC interval remains above the human positive prevalence of 0.119. On the held-out annotation split, AUROC is 0.717 and AUPRC is 0.113. The held-out AUPRC is 1.79 times the positive prevalence of $26/413 = 0.063$.

These ranking metrics use the continuous CLIP proxy distance, with human error as the positive class and equal weight per region over all 955 records. Recomputation gives AUROC 0.742787 and AUPRC 0.202825. Using the binary proxy label gives AUROC 0.659934. The confusion table uses binary labels; the ranking statistics use the distance before thresholding.

E.4 HUMAN-AUDITED POLICY UTILITY

Table 23(d) reports transfer of the original proxy threshold to the 36 held-out sources. DINO centroid yielded human risk 0.014 [0.000, 0.033], region coverage 0.657 [0.532, 0.770], and area coverage 0.664 [0.534, 0.787]. The separately human-calibrated row evaluates utility under its own calibration target.

Release-all yielded risk 0.0663 [0.0294, 0.1096] at full coverage, providing an uncalibrated descriptive reference for both policies. Under the same source aggregation and fixed candidate denominators,

Table 23: **Human-audit composition, annotation quality, and policy validation.** (a,b) Final counts and proxy alignment for all 955 regions. (c) Initial inter-rater and repeated intra-rater agreement. (d) Policy results on 36 held-out sources, with 95% source-bootstrap intervals. The two columns use their stated calibration targets; label alignment over all splits is not held-out policy performance.

(a) Source-disjoint audit split

Split	Sources	Items	Positive	Other
Dev	18	142	25	117
Calib	35	400	63	337
Test	36	413	26	387

(b) Proxy alignment with final human labels

	Proxy +	Proxy −	Total
Human positive	61	53	114
Other human labels	181	660	841
Total	242	713	955

(c) Inter- and intra-rater agreement

Comparison	Items	Agreement	Cohen's κ
01-02	319	0.909	0.498
01-03	318	0.950	0.627
02-03	318	0.921	0.600
Annotator 01	63	0.968	0.488
Annotator 02	63	0.968	0.871
Annotator 03	64	0.953	0.775

(d) Held-out policy validation

Metric	Proxy calibration	Human calibration
Threshold	167.759	568.481
Human risk	0.014 [0.000, 0.033]	0.057 [0.022, 0.097]
Region coverage	0.657 [0.532, 0.770]	0.951 [0.896, 0.990]
Area coverage	0.664 [0.534, 0.787]	0.952 [0.898, 0.991]
Conditional error	0.045	0.059

releasing a subset cannot increase the empirical error burden above release-all. Thus, every subset on this audited test set has a risk point estimate below 0.1. This observation alone is insufficient evidence of a scoring advantage. The release-all interval also extends above 0.1, so its point estimate should not be interpreted as a population risk guarantee. Figure 4 and Appendix H.2 evaluate coverage under separate human calibration and ranking at matched release counts, including LPIPS and random-policy comparisons at both quotas.

F Robustness and Applicability

The preceding sections evaluate the frozen protocol under matched conditions. The following experiments apply predefined perturbations and measure how matched recalibration changes the operating point. The concluding subsection connects these observations to the method's calibrated scope and future extensions.

F.1 Predefined perturbations

The predefined suite contains nine common-mode and nine release-only conditions: CIELAB shifts with $\Delta E \in \{5, 10, 20\}$, Gaussian blur with $\sigma \in \{1, 2, 4\}$, and horizontal translations of 2, 4, and 8



Figure 10: **Proxy risk and score changes under controlled perturbations.** (a) Clean-threshold proxy risk versus score change; marker area encodes action changes. (b) Clean-threshold proxy risk minus 0.10, with circles above the budget. (c) Matched-calibration region coverage for the perturbations and clean-threshold coverage for auxiliary collapse; circles mark no-release source rates above 0.50. Collapse repeats one auxiliary sample eight times and does not represent all mechanisms of diversity loss.

pixels. Transformations are deterministic and confined to the missing mask. Common-mode stress applies the same transformation to the designated and all eight auxiliary completions; release-only stress modifies only the designated completion. Each condition is evaluated with the clean threshold and with bias-matched calibration.

F.2 Aggregate behavior

The constructed scenario operating points have label risk 0.330 at coverage 0.635 and label risk 0.145 at coverage 0.315. Area coverage decreases from 0.467 to 0.231. These descriptive operating points do not establish a guarantee for each geometry subgroup. For DINO centroid under common-mode perturbations, 30 of 36 restorer–perturbation conditions exceeded proxy risk 0.1; none of the 36 release-only conditions did. Shared perturbations could lower disagreement while increasing erroneous releases, as shown in Figure 10.

Table 24 summarizes the clean-threshold and bias-matched behavior across the predefined perturbations.

At the frozen clean threshold, all 36 DINO centroid release-only conditions had observed proxy risk at most 0.1. Table 4 adds the direct four-score comparison; Table 25 separates all four restorers. For common-mode perturbations, matched calibration lowers mean proxy risk for every restorer (Table 24).

Table 24: **Stress summary averaged across the nine severities in each application family.** Exceed. counts clean-to-biased conditions with proxy risk above 0.1. The averages summarize responses to the controlled perturbation suite.

Restorer	Application	Exceed.	Clean proxy risk	Clean coverage	Score change	Matched proxy risk	Matched coverage	Matched no release
SD-inpaint	Common mode	8/9	0.2080	0.6810	-28.2	0.0850	0.4170	0.2610
SD-inpaint	Release only	0/9	0.0560	0.6620	199.3	0.0900	0.4800	0.1440
EAMamba	Common mode	4/9	0.1040	0.2150	-19.9	0.0790	0.1820	0.4890
EAMamba	Release only	0/9	0.0160	0.3430	213.3	0.0890	0.1650	0.4170
DDNM	Common mode	9/9	0.2090	0.6010	-10.1	0.1160	0.4410	0.3290
DDNM	Release only	0/9	0.0380	0.5210	241.9	0.1030	0.4110	0.1830
DPS	Common mode	9/9	0.2240	0.6410	-29.3	0.0970	0.4190	0.2760
DPS	Release only	0/9	0.0660	0.6870	187.6	0.0930	0.4460	0.1520

Table 25: **Release-only comparison by restorer and score.** Each row averages nine perturbation conditions. Frozen and matched columns use separate calibration settings at $\alpha = 0.1$. Conditional error averages sources with releases. These descriptive means do not imply significance.

Group	Score	Frozen clean threshold			Matched recalibration			
		Proxy risk	Cond. error (%)	Action change (%)	Proxy risk	Coverage (%)	No release (%)	AUROC
SD-inpaint	DINO centroid (primary)	0.0557	18.85	31.44	0.0898	47.99	14.43	0.7364
	DINO spread	0.1985	30.79	0.00	0.0955	43.73	23.26	0.6613
	DINO pair	0.0789	21.31	22.91	0.0937	49.16	14.83	0.7421
	Auxiliary-pair LPIPS	0.1737	29.85	0.00	0.1004	39.41	37.13	0.6011
EAMamba	DINO centroid (primary)	0.0164	31.12	11.13	0.0887	16.51	41.69	0.6450
	DINO spread	0.0690	48.79	0.00	0.0781	17.00	50.08	0.7460
	DINO pair	0.0216	34.12	8.93	0.0848	16.66	41.89	0.6754
	Auxiliary-pair LPIPS	0.0770	46.26	0.00	0.0850	17.75	63.59	0.7544
DDNM	DINO centroid (primary)	0.0380	17.68	36.90	0.1030	41.12	18.35	0.7202
	DINO spread	0.2067	34.65	0.00	0.1130	43.82	32.73	0.7151
	DINO pair	0.0583	20.90	27.78	0.1069	44.07	15.87	0.7465
	Auxiliary-pair LPIPS	0.1932	36.15	0.00	0.1097	41.21	40.33	0.6803
DPS	DINO centroid (primary)	0.0664	23.75	24.13	0.0925	44.62	15.23	0.7847
	DINO spread	0.1825	33.77	0.00	0.0953	41.52	28.34	0.7421
	DINO pair	0.0864	26.07	17.86	0.0937	45.56	15.23	0.7967
	Auxiliary-pair LPIPS	0.1831	33.91	0.00	0.0958	40.70	37.17	0.7248

F.3 Release-only perturbation families

Table 26 groups all nine release-only conditions into color, blur, and translation, retaining all four scores and both calibration settings. Auxiliary-only scores and their frozen-threshold actions are unchanged; matched recalibration can nevertheless change their thresholds because calibration labels change. The DINO pair mean is the mean squared distance from the designated output to the auxiliary completions. Equation (6) gives identical raw score increments for pair mean and the centroid score when only the designated output changes; their baseline scores and calibrated decisions can differ. The color results show a coverage advantage over auxiliary-only scores after recalibration, while DINO pair mean has higher mean AUROC than DINO centroid. Blur and translation must be considered separately when assessing ranking: a lower frozen-threshold proxy risk alone does not establish superiority at matched coverage. These controlled proxy-label comparisons do not measure historical fidelity or establish a human-risk guarantee.

Table 26: **Complete release-only perturbation-family comparison.** Each row averages 12 conditions: four restorers and three severities. Color uses $\Delta E = 5, 10, 20$; blur uses $\sigma = 1, 2, 4$; translation uses 2, 4, and 8 pixels. Conditions share sources and are not independent trials.

Group	Score	Frozen clean threshold			Matched recalibration			
		Proxy risk	Cond. error (%)	Action change (%)	Proxy risk	Coverage (%)	No release (%)	AUROC
Color	DINO centroid (primary)	0.0430	30.99	31.32	0.0925	28.27	31.86	0.7589
	DINO spread	0.2578	54.32	0.00	0.0922	21.72	53.48	0.6760
	DINO pair	0.0652	35.64	25.81	0.0924	28.23	31.68	0.7689
	Auxiliary-pair LPIPS	0.2526	55.60	0.00	0.0907	17.27	64.81	0.6195
Blur	DINO centroid (primary)	0.0431	16.00	22.91	0.0953	42.37	17.24	0.7322
	DINO spread	0.1163	28.50	0.00	0.0897	43.08	23.47	0.7651
	DINO pair	0.0586	19.11	15.98	0.0949	43.64	16.91	0.7523
	Auxiliary-pair LPIPS	0.1012	26.12	0.00	0.0955	44.27	33.48	0.7684
Translation	DINO centroid (primary)	0.0462	21.57	23.47	0.0928	42.04	18.17	0.6736
	DINO spread	0.1185	28.18	0.00	0.1046	44.75	23.86	0.7072
	DINO pair	0.0602	22.05	16.33	0.0970	44.70	17.27	0.6994
	Auxiliary-pair LPIPS	0.1165	27.91	0.00	0.1070	42.76	35.37	0.6826

Table 27: **Targeted stress diagnostics.** (a) Common CIELAB $\Delta E = 20$: clean-threshold and matched-calibration outcomes. (b) Auxiliary collapse: outcomes under the clean threshold only. Panels describe different perturbations and calibration protocols; their rows are not pooled.

(a) Shared color perturbation

Restorer	Clean proxy risk	Clean coverage	Score change	Matched proxy risk	Matched coverage	Matched no release
SD-inpaint	0.4080	0.6390	-19.4	0.0810	0.1280	0.6370
EAMamba	0.1480	0.1920	-21.1	0.0840	0.1150	0.6650
DDNM	0.4240	0.5930	-14.2	0.1070	0.1830	0.6260
DPS	0.4450	0.6080	-21.6	0.0720	0.1190	0.7160

(b) Auxiliary collapse

Restorer	Score change	Proxy risk	Region coverage	Action agreement	Release Jaccard	Flip rate
SD-inpaint	132.2	0.0560	0.4400	0.7820	0.6260	0.2180
EAMamba	137.7	0.0460	0.0990	0.9390	0.7860	0.0610
DDNM	77.7	0.0980	0.5220	0.9000	0.7810	0.1000
DPS	160.1	0.0430	0.4160	0.8170	0.6170	0.1830

F.4 Severe common color shift

Table 27(a) isolates the strongest common color shift in the predefined suite.

Under the strongest common color shift, matched calibration reduces proxy risk from 0.148–0.445 to 0.072–0.107 across the four restorers, with corresponding coverage changes reflecting adaptation to the shifted score distribution.

F.5 Auxiliary collapse

The controlled collapse diagnostic repeats auxiliary completion 01 eight times. Table 27(b) quantifies the resulting score and action stability.

Across all four restorers, proxy risk remained between 0.043 and 0.098, with action agreement of 0.782–0.939 under reduced auxiliary diversity.

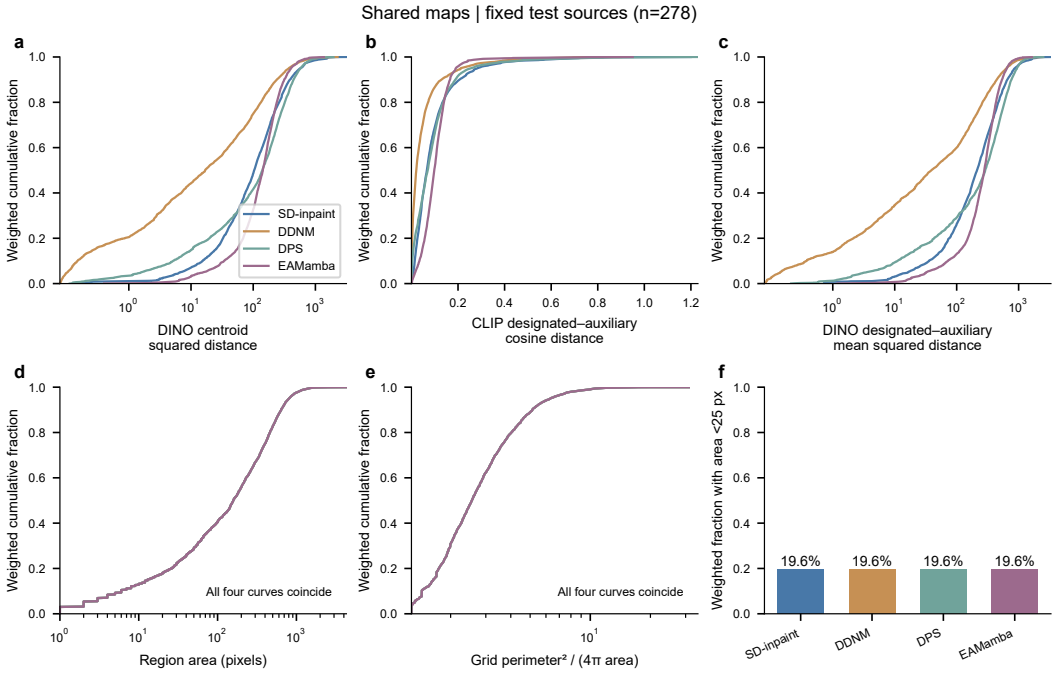

Figure 11: **Cross-restorer score and geometry distributions under Shared-map.** On 278 fixed test sources, panels (a–c) show DINO centroid, mean CLIP designated-to-auxiliary cosine distance, and mean squared DINO designated-to-auxiliary distance. Panels (d–f) show region area, grid perimeter complexity $P^2/(4\pi A)$, and the fraction below 25 pixels. Regions have source/instance-balanced weights $1/(278\,|\mathcal{I}_g|\,R_i)$, normalized over the displayed collection. Geometry distributions coincide because all restorers share the partition. DINO pair mean in (c) equals centroid plus auxiliary spread and is not independent evidence.

F.6 Scope and Future Extensions

Shared errors or limited spatial feature support can yield low disagreement even for incorrect regions. The formal guarantee of CSG applies to the proxy target, while the expert audit provides complementary diagnostics of construct validity and utility. Proxy risk within geometry groups and human error conditional on release remain elevated in the evaluation of the complete policy and tiny-region audit, respectively. We use $K = 8$ because larger banks change utility little in the evaluated settings. Future work may study efficient restoration schedules and evidence under shared distribution shifts.

G Diagnostic Analysis of EAMamba

EAMamba exhibits a distinct failure mode under both partition protocols. Under the shared spatial partition, DINO-centroid CSG releases 14.60% of candidate regions, with proxy risk 7.75%, conditional error 58.15% (95% CI: 51.37% to 64.83%), and AURC 0.3351. Restorer-native partitions yield 15.78% coverage, 8.19% proxy risk, conditional error 52.10% (95% CI: 45.14% to 58.97%), and AURC 0.3313. Native partitioning thus changes the operating point while preserving the low-coverage, high-error regime. The paired interval for the conditional-error change includes zero, so its observed reduction does not establish an improvement. These pointwise intervals use 5,000 source-bootstrap resamples conditional on the frozen outputs, split, labels, and calibrated thresholds; paired protocol intervals are reported in Appendix C.2.3.

The failure extends beyond tiny candidates. Among released regions of at least 25 pixels, the source/instance-balanced proxy-error proportions remain 54.73% under Shared-map and 58.35% under Native-map. These proportions retain each region's weight $1/(278\,|\mathcal{I}_g|\,R_i)$ from the full candidate collection, renormalized within the released size subgroup. They are not pixel-area-

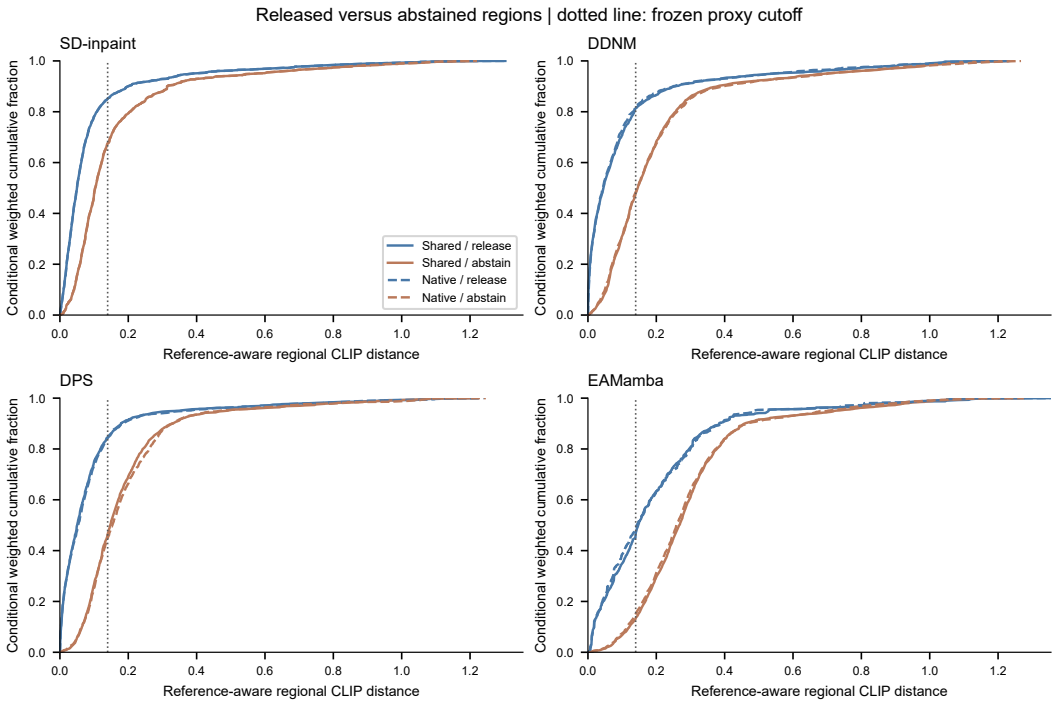

Figure 12: **Released EAMamba regions retain substantial reference-proxy error.** Curves show continuous reference-relative CLIP distances for released and abstained candidates under both protocols; dotted lines mark the frozen proxy cutoff. Source/instance-balanced regional weights are renormalized within each decision group. These descriptive conditional distributions use a different aggregation from source-macro conditional error. CLIP distances are reference-relative proxies, not human judgments or factual-error labels.

weighted and differ from the source-macro conditional error above. They provide a complementary diagnostic, not directly comparable estimates or a subgroup risk guarantee. Figure 13 illustrates a released proxy error outside the tiny-region group.

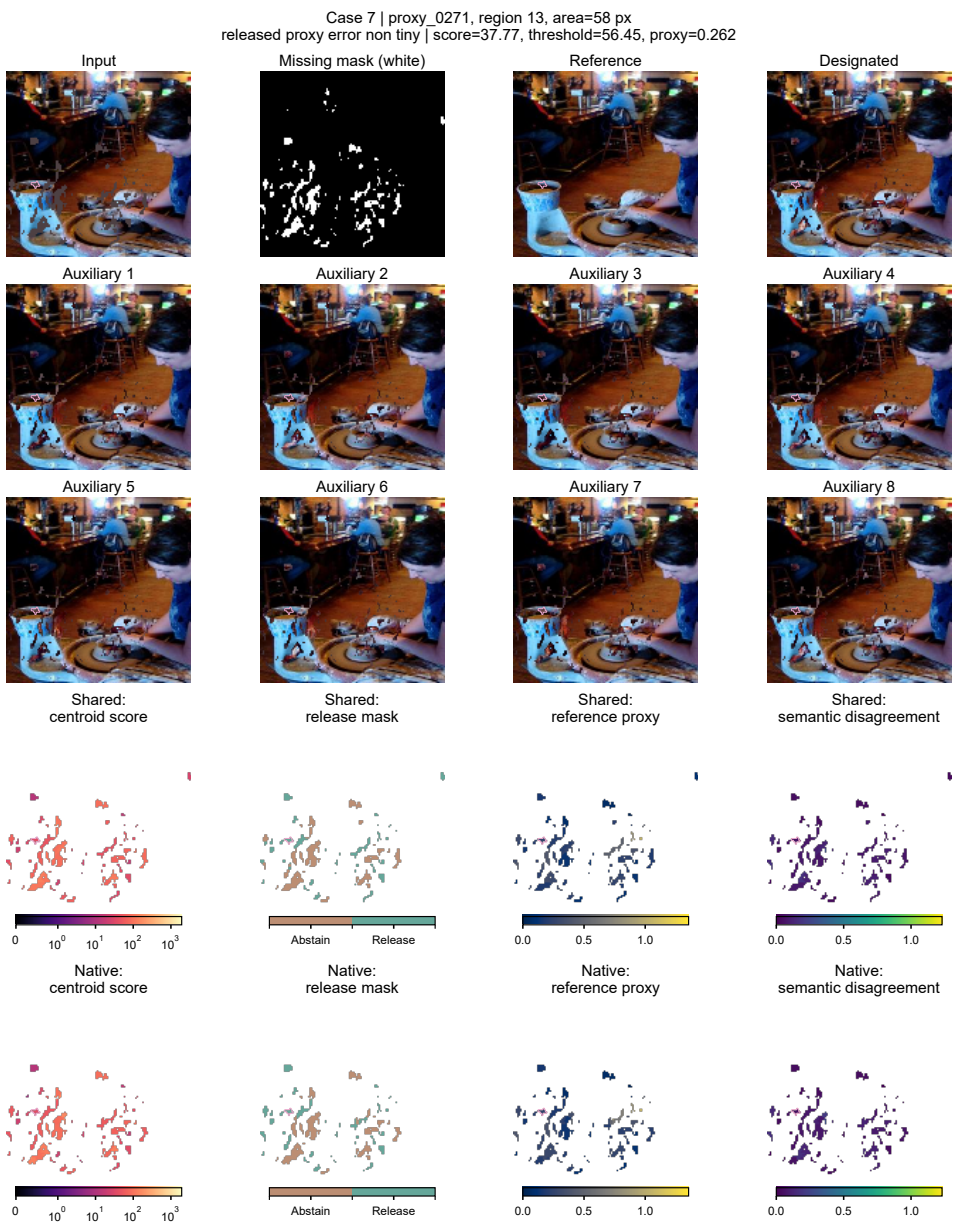

Figure 13: **A released proxy-positive EAMamba region larger than 25 pixels.** The shared-map candidate in instance `proxy_0271` has area 58 pixels, centroid score 37.77 below threshold 56.45, and reference-proxy distance 0.262. The plate retains the input, mask, reference, designated output, all eight auxiliaries, and both protocols' regional maps. "Semantic disagreement" denotes mean CLIP designated-to-auxiliary cosine distance; neither map is pixel-level ground truth. This illustrative case was selected near the median of a predefined released, proxy-positive, non-tiny group; it does not estimate failure frequency. Original restoration pixels are unchanged.

Score distributions indicate a dissociation between sample agreement and reference-relative correctness (Figures 11 and 12). Under Shared-map, EAMamba and DPS have similar weighted median DINO-centroid scores (142.43 and 140.47), whereas EAMamba's median reference-relative CLIP distance is larger (0.2528 versus 0.0869). Low designated-to-auxiliary disagreement therefore does not reliably imply reference agreement in this setting. This does not mean that all EAMamba samples have low disagreement. The analysis identifies an empirical limitation of the score instantiation on these cached outputs; it does not isolate training, sampling behavior, or implementation details as the cause.

Table 28: **Encoder–score comparison under proxy and human calibration.** (a) Means over 200 paired source re-splits, calibrated to the proxy at $\alpha = 0.1$. Conditional error averages over sources with releases; exceedance is the observed test-risk fraction above 0.1. (b) Separate human-label calibration and a fixed 36-source human test set. The panels have different label populations and cannot be interpreted as paired proxy-versus-human risk estimates. Ranks compare deployable scores within each panel, excluding the reference-aware oracle in (a); risk and exceedance are calibration diagnostics.

(a) Proxy-calibrated re-split results

Score	Risk	Reg. cov. ↑	Area coverage ↑	Cond. error ↓	AURC ↓	Exceed.
DINO centroid (primary)	0.098	0.653	0.639	0.162	0.069	0.435
DINO spread	0.097	0.651	0.642	0.162	0.069	0.400
DINO pair	0.097	0.667	0.656	0.157	0.067	0.440
CLIP centroid	0.097	**0.772**	**0.783**	**0.145**	**0.049**	0.395
CLIP spread	0.097	0.734	0.744	0.155	0.055	0.380
CLIP pair	0.097	0.766	0.775	0.148	**0.049**	0.375
Random	0.096	0.452	0.451	0.223	0.098	0.340
Reference-aware oracle	0.097	0.846	0.843	0.136	0.021	0.415

(b) Separately human-calibrated results

Score	Human risk	Human cov. ↑
DINO centroid (primary)	0.057	**0.951**
DINO spread	0.057	0.870
DINO pair	0.054	0.929
CLIP centroid	0.049	0.889
CLIP spread	0.059	0.845
CLIP pair	0.057	0.859
LPIPS	0.058	0.941

H Supplemental Score, Geometry, and Human Audits

H.1 Score Evaluation Protocol

The six-score analysis uses all 1,000 natural-mask instances from 695 sources, including 51 instances with no candidate region, and reproduces the three previously available scores exactly. Development, calibration, and test sources remain separate. Each of the 200 re-splits uses a shared source assignment across methods, a fixed development grid, and a newly calibrated threshold for each score. Fixed-test intervals use 5,000 paired source-bootstrap samples conditional on the calibrated threshold. The empirical ranges describe variability across 200 re-splits of one source collection. Figure 3 displays region coverage versus conditional proxy error for all 13 scores and diagnostics from the original paired comparison. Each method contributes 200 observations, giving 2,600 distinct method–split observations. Repeating DINO centroid in all six panels gives 3,600 plotted points. The other 12 methods are randomly paired for display, without selecting observations. Marginal densities use Gaussian kernels with Scott bandwidths in the original percentage units. Panel-specific linear limits retain all observations; comparisons across panels should use the reported axis values. Table 28(a) separately reports the two-encoder, three-score comparison.

DINOv2-small outputs its final LayerNorm patch tokens on an 18×18 grid after resizing to 252×252. CLIP ViT-B/16 outputs the vision encoder patch tokens on a 14×14 grid at 224×224; these do not use the post-LayerNorm applied only to the pooled CLS token or the image-text projection. Both use their specified input normalization. Dense feature maps are linearly interpolated to 256×256 and region-averaged before squared distances, with no unit-L2 normalization. On natural test designated regions, median pooled norms are 42.244 for DINO and 13.826 for CLIP. These are not distances between unit vectors. The re-extraction uses FP32 inference with transformers 5.4.0 and reproduces the previous scores numerically.

H.2 Human Calibration and Matched Ranking

The 955-region audit contains 18 development, 35 calibration, and 36 test sources, including 413 test regions and 26 identified errors. Eligible regions contain at least 25 pixels. All nine initial uncertain judgments were adjudicated, yielding five positive and four negative labels. Human calibration fits a separate threshold using calibration labels, while proxy-threshold transfer retains the original threshold.

Utility of separately calibrated policies. DINO centroid attained region coverage 95.1% at risk 0.0566. DINO spread attained 87.0%/0.0570, CLIP spread 84.5%/0.0587, and LPIPS 94.1%/0.0578. These observed operating points describe coverage and risk under each score's fitted threshold. Release-all yielded risk 0.0663 with a 95% source-bootstrap interval of [0.0294, 0.1096] at full coverage. It provides an uncalibrated descriptive reference. With fixed denominators, all release subsets have empirical risk bounded by this release-all point estimate; being below 0.1 does not demonstrate a scoring advantage.

Ranking at two release quotas. Each score releases the same count within every audited instance as DINO centroid at the specified threshold, using score rank with ties broken by region ID. The low quotas come from proxy-calibrated DINO centroid and yield 257 releases, corresponding to 65.7% coverage averaged over sources. DINO centroid risk was 0.01404 and LPIPS risk was 0.01628, with LPIPS marginal 95% interval [0.00000, 0.03667]. The DINO centroid-minus-LPIPS point difference was -0.00223. A paired interval is not reported for the corrected low-quota comparison. The random-policy mean was 0.0168 and its central 95% randomization range was [0.0062, 0.0267].

The high quotas come from human-calibrated DINO centroid and yield 391 releases, corresponding to 95.1% coverage. DINO centroid and LPIPS risks were 0.05659 and 0.05674, giving a paired DINO centroid-minus-LPIPS difference of -0.00015 [-0.00463, 0.00417]. The separate random-policy mean and range at these quotas were 0.0613 and [0.0567, 0.0647]. The paired interval at the higher quota includes zero, and the lower quota comparison lacks a paired interval.

Conditional uncertainty. Marginal and paired intervals use 5,000 source-bootstrap replicates; paired differences use the same resampled source identities for both methods. Point differences are computed from the observed estimates. These intervals condition on the current labels, source split, fitted thresholds, and per-instance quotas. They do not include uncertainty from recalibration or reannotation. Randomization ranges use 1,000 random policies computed separately for each quota on the observed test sample and describe variation in random selection. Counts are matched within instances; reported coverage averages within sources and then across sources, so it need not equal the pooled released-region fraction. Figure 4 displays these distinct comparisons.

H.3 Geometry Interventions and Tiny-Region Audit

For original candidate regions r, the descriptive area burden is $\sum_r A_r y_r I_r / \sum_r A_r$, with original region area A_r, proxy label y_r, and release indicator I_r. It is averaged within source and then across sources. Rejected tiny regions remain in the denominator. Neighborhood pooling gathers regions in the same connected candidate component by centroid distance, breaking ties by region ID, until 25 pixels are supported or the component is exhausted. Features weighted by area are used for each original decision region; neighborhoods may overlap and original labels are retained. This intervention changes feature pooling while preserving the original regions. Compared with original pooling, its paired overall area-burden difference is -0.00957 (95% interval $[-0.01262, -0.00677]$), while area coverage changes by -0.01221 ($[-0.01608, -0.00863]$); the proxy-risk difference interval includes zero.

The additional tiny-region audit retains 0/1/U responses and uses a third reviewer for any primary disagreement or U. Original agreement is computed before adjudication. Reviewers see reference and designated images plus same-location nearest-neighbor enlargements; surrounding content is context, not the target being labeled. One-pixel regions can be unassessable even when enlarged. The primary analysis excludes T153, the assisted item, leaving 199 sources with 110 no-error labels, 88 error labels, and one unresolved label. The frozen policy releases 30 regions, including 19 errors; the conditional error proportion is 0.6333 with a 95% interval of [0.4516, 0.8000] from 5,000 bootstrap

Table 29: **Geometry interventions and the tiny-region audit.** (a) Mitigation on the original candidate support: All denotes 320 test sources; Thin denotes 80 long thin sources. Area error is proxy-weighted released area, averaged over sources. (b) Separate audit of one selected region below 25 pixels per source, 199 sources after excluding T153. U denotes unresolved adjudication. These panels use different populations and denominators. Ranks compare interventions separately within All and Thin in (a); counts in (b) are descriptive.

(a) Risk, coverage, and area burden after intervention

Setting	Risk	Region coverage ↑	Area coverage ↑	Area error ↓
All: Original	0.104	0.151	0.173	0.110
All: Guard 25	0.105	**0.154**	**0.278**	0.195
All: Pool 25	0.104	0.151	0.161	**0.101**
Thin: Original	0.417	**0.601**	0.685	0.442
Thin: Guard 25	0.343	0.474	**0.940**	0.685
Thin: Pool 25	0.417	0.599	0.639	**0.404**

(b) Final tiny-region labels by geometry

Geometry	Sources	No error	Error	U
Boundary intersecting	50	26	23	1
Compact irregular	50	36	14	0
Long thin	49	17	32	0
Multiple components	50	31	19	0

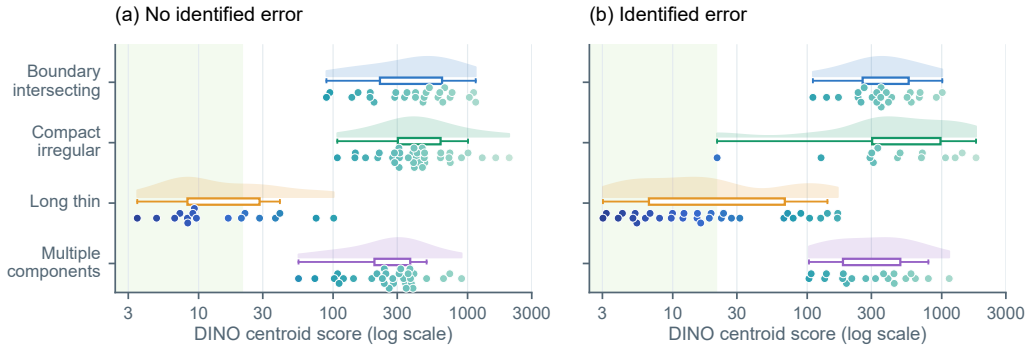

Figure 14: **Tiny-region scores by human label and geometry.** (a) 110 regions with no identified error after excluding T153; (b) 88 regions with an identified error. Each point represents one region from a distinct source. Lower scores are darker on a shared logarithmic scale. Pale green marks scores at or below the frozen threshold, 20.924512. Boxes show medians and quartiles with 1.5-interquartile-range whiskers. Half-densities use $\log_{10}$ scores, a common bandwidth of 0.18, observed-range support, and normalized peaks; their colors identify geometry. The unresolved item is retained in Table 29(b). The selected audit does not estimate population-wide error prevalence.

resamples stratified by geometry. Each source contributes one selected region. These intervals condition on the selected records, labels and threshold and do not estimate deployment prevalence. The sensitivity analysis including T153 gives 31 releases and 19 errors, or 0.6129. Figure 14 uses the primary exclusion analysis: 11 released regions have no identified error and 19 have an identified error. The unresolved item is withheld and remains outside the plotted label groups.

H.4 Runtime Measurements

On an RTX 5080, the four-restorer cached-input DINO centroid processing benchmark has 1,600 observations (100 timed sources per restorer and K setting after 20 warmup sources). At $K = 8$,

Table 30: **Region-count versus candidate-area calibration.** Fixed-test source averages at $\alpha = 0.1$; each restorer uses the same scores and source split for both losses. Area metrics use fixed candidate area. Conditional error is averaged over sources with releases.

Restorer	Loss	Region risk	Region coverage	Area-error burden	Area coverage	Conditional error
SD-inpaint	Region count	0.0927	0.6272	0.0912	0.6105	0.1510
SD-inpaint	Candidate area	0.0970	0.6521	0.0961	0.6383	0.1515
EAMamba	Region count	0.0775	0.1460	0.0544	0.1059	0.5815
EAMamba	Candidate area	0.0966	0.1704	0.0719	0.1282	0.5985
DDNM	Region count	0.1085	0.5810	0.1184	0.5558	0.1889
DDNM	Candidate area	0.0997	0.5504	0.1094	0.5266	0.1877
DPS	Region count	0.0891	0.5801	0.0887	0.5491	0.2033
DPS	Candidate area	0.0988	0.6133	0.0976	0.5798	0.2067

mean processing time is 0.460–0.474 seconds; it excludes generation, disk input, and image decoding. Three of the 100 timing-only region reconstructions differ from their historical boundaries; these timing runs are excluded from accuracy comparisons. Separate SD measurements contain 2,040 generations, including warmup, at 50 steps, native 512×512, FP16 and batch size one, with output projected to 256×256. At $K = 8$, mean generation time for one designated and eight auxiliary outputs is 17.067 seconds. The separately measured component quantiles cannot be added to obtain end-to-end p95. Formal generation timings for EAMamba, DDNM, and DPS remain unavailable because the compiled environment or identified checkpoints are missing. The reported timings cover separate components, not matched end-to-end execution across all four models.

I Additional Loss and Score Comparisons

I.1 Candidate-area Calibration

Region-count burden gives equal weight to erroneous released candidate regions. It addresses how much regional error is exposed from the fixed support, allowing a policy to abstain. It does not measure precision among released regions, pixelwise correctness, or semantic severity. For one instance with at least one release, its region-count burden equals its region coverage times its error conditional on release. This identity need not hold after separately averaging each metric over sources, particularly because conditional error excludes sources without releases.

An alternative target weights each regional proxy label by its fixed area:

$$\ell_i^{\mathrm{area}}(\lambda) = \begin{cases} \dfrac{\sum_{r \in \mathcal{R}_i} A_{ir} \tilde{h}_{ir} \mathbf{1}\{s_{ir} \le \lambda\}}{\sum_{r \in \mathcal{R}_i} A_{ir}}, & R_i > 0, \\ 0, & R_i = 0. \end{cases} \tag{7}$$

Areas and proxy labels are fixed before threshold selection. The loss is bounded in $[0, 1]$ and nondecreasing in λ; averaging within source and applying Equation (4) gives the same marginal expected-loss argument under the frozen protocol and source exchangeability. The denominator is candidate area, including abstained candidates, and the labels remain regional proxy errors. This regional error burden weights proxy labels by area.

Table 30 uses the same DINO centroid scores, development grids, source split, candidate maps, and proxy labels for both objectives, with $\alpha = 0.1$. Only the calibration loss changes. SD-inpaint, EAMamba, and DPS gain coverage under area calibration; DDNM loses coverage and reduces area burden from 0.1184 to 0.1094. Its latter fixed-test estimate still exceeds 0.1. The guarantee concerns marginal expected loss. EAMamba retains high conditional error under both objectives. Objective choice changes the release tradeoff without ensuring precision or resolving restorer-specific failure modes.

I.2 Exploratory Human Budget Sweep

We use the existing 18/35/36 development/calibration/test human sources, the same 413 labeled test regions, and each score's development-fitted grid. The budgets $\{0.01, 0.025, 0.05\}$ were requested

Table 31: **Exploratory human calibration at tighter budgets.** Each method uses its own development score grid and the same 35 calibration and 36 test sources. Intervals are marginal 95% source-bootstrap intervals from 5,000 draws, conditional on the labels, split, and fitted threshold. LPIPS denotes auxiliary-pair LPIPS.

Budget	Score	Human risk [95% interval]	Coverage [95% interval]
0.01, 0.025	All four scores	0 (all abstain)	0 (all abstain)
0.05	DINO centroid	0.0140 [0.0000, 0.0335]	0.6900 [0.5689, 0.8029]
0.05	DINO pair	0.0158 [0.0015, 0.0351]	0.7058 [0.5849, 0.8140]
0.05	LPIPS	0.0229 [0.0000, 0.0562]	0.7987 [0.6822, 0.9000]
0.05	Fixed random score	0.0169 [0.0029, 0.0363]	0.2680 [0.2071, 0.3402]

Table 32: **Designated-to-auxiliary LPIPS under proxy calibration.** (a) Clean fixed-test source averages with separate thresholds at $\alpha = 0.1$. (b) Equal-weight means over all 36 release-only conditions, plus the 12-condition color subset. Exceedance counts describe test estimates above 0.1; the common source split makes conditions dependent. RC: region coverage; CE: error conditional on release; NR: no-release source rate.

(a) Clean operating points

Restorer	Risk	RC	Area coverage	CE
SD-inpaint	0.10000	0.58891	0.54514	0.15585
EAMamba	0.08693	0.14971	0.12526	0.54574
DDNM	0.11122	0.55713	0.53291	0.21352
DPS	0.09672	0.59832	0.55246	0.20459

(b) Release-only perturbations

Family	Threshold	Risk	Max. risk	RC	CE	NR	> 0.1
All	Clean frozen	0.0842	0.2164	0.3259	0.2998	0.4246	11/36
All	Matched	0.0973	0.1327	0.3636	0.3321	0.3694	17/36
Color	Clean frozen	0.1109	0.2164	0.2748	0.4268	0.4850	6/12
Color	Matched	0.0878	0.1051	0.2161	0.4384	0.5507	1/12

after inspection of the original human study; this is an exploratory sensitivity analysis, with no selection of a preferred policy using test labels. Random scoring uses the existing fixed random scores, not a newly selected favorable seed.

With $G = 35$ calibration sources, the correction in Equation (4) is at least $1/(G+1) = 1/36 \approx 0.0278$. Consequently no grid threshold is feasible at 0.01 or 0.025, and the protocol returns the explicit all-abstain action for every score. At 0.05, LPIPS has higher observed coverage than DINO centroid and DINO pair, at higher observed risk; all three remain below 0.05 on this test set. Coverage and human risk increase together at this budget; the result applies to the observed test split.

I.3 Designated-to-auxiliary LPIPS

The added score in Appendix B.5 retains LPIPS-Alex and regional pooling while comparing the deliverable with each auxiliary. We evaluate all 1,000 cached instances for each of four restorers, on clean images and all nine predefined release-only perturbations. Candidate maps, auxiliary images, and regional proxy labels follow the original protocol. Clean and matched thresholds use their respective development grids and calibration sources; no new restored images or human labels are collected.

With clean thresholds, designated LPIPS changes 15.3% of regional decisions on average across 36 conditions, at risk 0.0842 and coverage 32.6%. Risk exceeds 0.1 in 11/36 conditions. Matched recalibration gives risk 0.0973, coverage 36.4%, and no-release source rate 36.9%. The color subset has matched risk 0.0878 and coverage 21.6%. Designated LPIPS responds when the deliverable changes because the deliverable enters the score directly. Its remaining risk exceedances show that access to that output does not alone ensure robustness under shift.

Table 33: **Human evaluation of designated-to-auxiliary LPIPS.** (a) Separate human calibration at $\alpha = 0.1$ on the same 35 sources. (b) The same low/high instance quotas as Figure 4. Intervals use 5,000 paired source-bootstrap draws on 36 test sources, conditional on labels, split, thresholds, and quotas. Difference is observed DINO centroid minus designated LPIPS risk. All quantities are fractions.

(a) Separately calibrated utility

Score	Human risk	Region coverage
DINO centroid	0.05659	0.95051
Auxiliary-pair LPIPS	0.05779	0.94071
Designated-to-auxiliary LPIPS	0.05987	0.96096

(b) Ranking at matched instance quotas

Quota	Releases	Centroid risk	LPIPS-del risk	Difference [95% interval]
Low	257	0.01404	0.00139	0.01265 [-0.00139, 0.03210]
High	391	0.05659	0.05848	-0.00189 [-0.00772, 0.00278]

Under separate human calibration, designated LPIPS attains risk 0.05987 and coverage 96.1%, compared with 0.05659/95.1% for DINO centroid. At 257 matched releases its risk is 0.00139; the observed centroid-minus-LPIPS difference is 0.01265 with paired interval $[-0.00139, 0.03210]$. At 391 releases its risk is 0.05848; the difference is -0.00189 with interval $[-0.00772, 0.00278]$. Both intervals include zero. These comparisons retain the distinction between calibrated policy utility and ranking at matched counts, and do not support uniform ranking dominance of the centroid score.

J Additional checks of the calibration target and evaluation unit

J.1 Target population and deployment unit

A source group contains an image identity, its instance count N_g, its masks, and the retained generation randomness. The guarantee assumes that calibration and future groups follow the same frozen joint mechanism. Group sizes may vary. It controls $G^{-1}\sum_g L_g$ in expectation over future groups, while a stream that weights every instance equally targets $\sum_g N_g L_g / \sum_g N_g$. For a group drawn from the target population and an instance J drawn uniformly within that group, $\mathbb{E}[\ell_{gJ}(\hat{\lambda}) \mid g] = L_g(\hat{\lambda})$. This identity does not extend the guarantee to an arbitrary request stream. The archived benchmark supplies a finite image-and-mask collection; exchangeability of a deployment population is an assumption, not a property established by these tests.

On the original fixed test split, DINO centroid has source-mean proxy risk 0.0927 and instance-weighted proxy risk 0.0923; the corresponding region coverages are 0.6272 and 0.6155. An exploratory check selects one instance uniformly within each source, independently of scores and labels, then refits the development score grid and calibration threshold using the existing source partitions. The original proxy target is retained. This policy attains risk 0.0918 and region coverage 0.6266 on 278 test sources. The selection, per-method results, and paired source intervals are supplied in the revision analysis files. This check does not establish exchangeability of a real request stream.

J.2 Metric definitions used by the cached analyses

For an instance with candidate areas a_{ir} and release indicators I_{ir}, the primary cached analyses compute candidate-area coverage as $\sum_r a_{ir} I_{ir} / \sum_r a_{ir}$. Coverage of the entire missing mask would instead divide by $|M_i|$. These denominators can differ because some missing pixels have no retained candidate region. Candidate-area coverage is averaged within instances and then sources. Empty candidate instances contribute zero. The geometry benchmark has its separately specified area summaries; these should not be equated to the primary candidate-area metric without checking denominators.

Table 34: Exploratory proxy-target sensitivity for the primary DINO centroid score. All rows use the original 278-source test split and $\alpha = 0.10$. Coverage is regional coverage. The release-all reference uses the same labels and source aggregation. Full four-score results are supplied with the analysis files.

q	τ_q	Proxy risk	Coverage	No-release sources	Release-all risk
0.60	0.092287	0.0869	0.4305	0.1763	0.3357
0.70	0.119748	0.0942	0.5565	0.1223	0.2414
0.75	0.139669	0.0927	0.6272	0.0791	0.1962
0.80	0.176282	0.0946	0.7715	0.0504	0.1506
0.90	0.336662	0.0676	0.9460	0.0324	0.0676

Conditional error first pools released errors and release counts within a source, then averages their ratio across sources with at least one release. It is not the mean of instance-specific ratios. Primary AURC integrates the fixed-denominator proxy burden against source-mean region coverage by the trapezoidal rule, on the unique development quantile grid augmented by both infinite sentinels. Coverage is stably sorted; repeated coverage points have zero width. No normalization by the maximum attainable coverage is applied. The all-release endpoint can be below one because empty candidate instances have zero coverage. The stress analysis uses its own recorded curve grid and should not be compared numerically to the primary AURC without matching these definitions.

J.3 Proxy-target sensitivity

The proxy threshold τ determines which regions count as erroneous; α budgets their released burden; λ selects the release set. The main $\tau = 0.13966859463872094$ is the 0.75 quantile of the development CLIP distances with equal total weight per source. It is not an application-validated distortion limit. The exploratory sweep below fixes outputs and scores, uses the same source-balanced quantile rule, recomputes proxy labels and recalibrates each score. The five quantiles were chosen during this revision. Changing τ changes the error task; higher coverage under a more permissive label is not a method gain.

J.4 Result versions and remaining evidence limits

The original budget sweep and the primary score comparison use the same pool of 556 non-development sources but different random split sequences. No corresponding test-source set matches across their 200 repetitions. This revision recomputes all budget points using the primary split sequence; at $\alpha = 0.10$, its risk is 0.097572 and its exceedance frequency is 0.435. Figure 2(a) and Table 8(a) now use that sequence. These are empirical split ranges on a fixed source collection, not confidence intervals based on 200 independent datasets.

Reaggregation of the 36 release-only conditions gives matched proxy risk 0.093519. The three matched family risks are 0.092501, 0.095285 and 0.092772 for color, blur and translation. Tables reporting these matched-risk values use this condition table. The earlier manually supplied value 0.0895 is not reproduced by that archive.

The corrected low-quota auxiliary LPIPS point estimate was supplied without a replacement regional release mask or source loss vector. Its paired interval cannot be reconstructed from marginal intervals; no estimated paired interval is reported. Source overlap tables establish recorded partition membership, not historical independence from all protocol development or qualitative selection. Historical generation caches also do not establish end-to-end reproducibility without their checkpoints, complete configurations and output-role manifests.

Auxiliary-bank aggregation. The archived stability implementation aligns regions by source, instance and region ID and compares each bank with the primary bank. Spearman correlation uses all aligned region scores pooled together, with average ranks for ties; constant vectors return an undefined correlation under the SciPy routine. The primary bank compared with itself is assigned correlation one. Action agreement, flips and Jaccard are computed within source and then averaged; an empty release union has Jaccard one. These aggregation units differ and are reported separately.

Primary and replicate seed plans are included in the revision analysis inventory; the plans do not alone verify the availability or identity of all historical image outputs.