RIMAR: RIEMANNIAN METRIC ATTENTION AND REGULARIZATION FOR TRAINING-FREE OPEN-VOCABULARY SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free open-vocabulary semantic segmentation converts frozen vision-language features into pixel predictions. Existing inference methods typically use different geometric quantities for feature comparison and spatial refinement. We introduce RIMAR, which uses a Riemannian metric derived from spatial variation in a multi-layer feature field as a common geometry for both. Bidirectional Affine Attention (BAFAT) fits closed-form affine query-key mappings using normalized metric-area weights. Its area-weighted proposal is fused with native attention before value aggregation. The retained metric supplies directional distances to a refinement graph alongside RGB differences. Semantic information from the input and overlapping-window integration produce observations refined by fixed-budget Gaussian and heavy-tailed updates derived from maximum a posteriori (MAP) objectives. RIMAR requires no segmentation training, auxiliary visual backbone, or camera calibration. For a continuous BAFAT-MAP reference family, we establish reparameterization equivariance of scores and MAP solution sets under consistent transport of features, sampling measures, and regularization fields. With frozen CLIP ViT-B/16, RIMAR averages 41.4% mIoU and 70.6% pixel accuracy across 13 panoramic, fisheye, and perspective benchmark settings, exceeding PEARL by 2.9 and 3.8 percentage points, respectively, under a common evaluation protocol. Paired projection and sampling studies show higher mean accuracy and prediction agreement than PEARL and NACLIP, alongside residual sensitivity. Controlled comparisons show an upstream advantage over PEARL under a common refiner and a 0.44-point mIoU improvement over a jointly simplified geometry-independent variant.

1 INTRODUCTION

Training-free open-vocabulary semantic segmentation (OVSS) repurposes a frozen vision-language model for dense prediction without segmentation-specific training. A user supplies the target categories as text, allowing the vocabulary to change without updating the model. CLIP provides the image-text alignment needed for this transfer (Radford et al., 2021), but recognizing a concept somewhere in an image does not ensure precise localization of its constituent pixels. The challenge is to recover spatially precise predictions while preserving the pretrained model's semantic flexibility.

Recent progress demonstrates the importance of controlling how frozen features are aggregated. GEM and SCLIP reorganize token interactions to expose local structure (Bousselham et al., 2024; Wang et al., 2024), while ClearCLIP identifies globally dominant representation components that obscure dense predictions (Lan et al., 2024a). NACLIP encourages neighborhood awareness (Hajimiri et al., 2025), and SFP suppresses outlier propagation (Jin et al., 2025). ResCLIP combines intermediate-layer attention with semantic feedback (Yang et al., 2025), while CDAM constructs affinities from patch-level class distributions (Kang et al., 2025). These results show that dense inference depends not only on individual patch representations, but also on how information is exchanged between them.

The same issue extends to broader context and spatial refinement. GLA-CLIP and OV-Stitcher enable attention across otherwise independently processed windows (Lee et al., 2026; Moon et al., 2026).

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.

NERVE combines entropy-based attention fusion with random-walk propagation using diffusion-model neighborhoods (Mahatha et al., 2026). PEARL provides a particularly close reference by combining orthogonal key-query alignment with text-aware Laplacian propagation (Pei et al., 2026).

These advances motivate a more specific question: can patch aggregation and score refinement derive their spatial quantities from the same feature geometry? A spatial metric provides both a local area measure for weighting patch contributions and directional lengths for defining spatial coupling. We investigate this connection across perspective, panoramic, and fisheye imagery, deriving geometry from feature variation without a camera model and retaining RGB differences as a separate appearance cue.

RIMAR realizes this construction through a shared feature-induced Riemannian metric (Fig. 1). Spatial variation of a standardized multi-layer patch field defines the metric. Bidirectional Affine Attention (BAFAT) uses its normalized area weights for closed-form affine query-key fits and the resulting attention proposal; its directional lengths enter the spatial refinement graph. The per-view metric is carried forward through window integration and resampling, so refinement uses the geometry retained from attention. The complete predictor combines native attention with BAFAT, adapts temporary text representations, and integrates local and contextual scores across windows. Confidence-weighted Gaussian refinement, an original-resolution blend, and fixed-budget heavy-tailed updates produce the final predictions. Both encoders remain frozen, and no auxiliary visual backbone is required.

We establish reparameterization equivariance of semantic scores and maximum a posteriori (MAP) solution sets for a continuous BAFAT-MAP reference family under consistent field transport. This result concerns the reference construction rather than exact equivariance of the finite image-to-label predictor. With frozen CLIP ViT-B/16, RIMAR obtains 41.4% mean mIoU and 70.6% pixel accuracy across 13 benchmark settings, exceeding PEARL by 2.9 and 3.8 percentage points, respectively, under identical evaluation conditions. Controlled comparisons distinguish complete-system gains from the incremental contributions of particular aggregation and refinement choices.

Our contributions are:

- a shared feature-geometric construction whose area measure weights bidirectional affine attention and whose directional lengths guide spatial score refinement within a frozen-model predictor;
- a reparameterization-equivariance result for scores and MAP solution sets of a continuous BAFAT-MAP reference family under explicit transport assumptions; and
- an evaluation across 13 perspective, panoramic, and fisheye benchmark settings, with controlled comparisons separating upstream aggregation, refinement, and the roles of geometric and RGB cues.

2 Related Work

Dense features from frozen vision-language models. Training-free methods recover dense structure by reorganizing token interactions: GEM and SCLIP revise attention, while NACLIP adds neighborhood awareness (Bousselham et al., 2024; Wang et al., 2024; Hajimiri et al., 2025). ClearCLIP and SFP address complementary sources of localization error by suppressing globally dominant components and outlier propagation, respectively (Lan et al., 2024a; Jin et al., 2025). CLIPtrase recovers inter-patch semantic correlations and uses clustering and denoising to form coherent regions (Shao et al., 2024). RIMAR draws on these distributed signals to construct its feature field and combine native attention with a metric-weighted affine proposal.

Semantic context and spatial propagation. Semantic ambiguity motivates class purification, class-distribution affinities, recurrent inference, and region graphs (Chen et al., 2025; Kang et al., 2025; Sun et al., 2024; Huang et al., 2026). PEARL provides the closest alignment-and-propagation reference (Pei et al., 2026), while GLA-CLIP emphasizes context across windows (Lee et al., 2026). RIMAR retains a feature-derived metric across attention and refinement, alongside semantic and window integration. Our common-refiner comparison with PEARL separates differences in the upstream semantic observations from differences in downstream propagation. Unlike systems that learn aggregation or add proxy encoders (Cho et al., 2024; Lan et al., 2024b; Wang et al., 2025),

it uses one frozen CLIP image/text pair; its image-conditioned prototypes are temporary inference tensors.

Geometry. Camera-specific spherical models encode known projection domains (Carlsson et al., 2024; Benny and Wolf, 2025), while RIMAR derives a pullback metric from spatial feature variation without a camera model instead (Lee, 2018). Its coordinate-consistency result assumes joint transport of continuous fields; independently rendered images and extracted features need not satisfy those assumptions.

3 METHOD

RIMAR connects patch aggregation and spatial score refinement through a shared feature-derived geometry (Fig. 1). For each image or crop, frozen CLIP supplies hidden states $S^{(\ell)}$, $\ell = 1, \dots, M$, native attention proposals $A^{(\ell)}$, and text prototypes t_a for A aliases assigned to semantic classes by π. Stage A constructs a patch field and semantic observations; Stage B refines them using the retained metric and a separate RGB cue. Here, a layer is a transformer-block position, not scene depth. Algorithm 1 and Appendix A give the complete computation and supporting definitions.

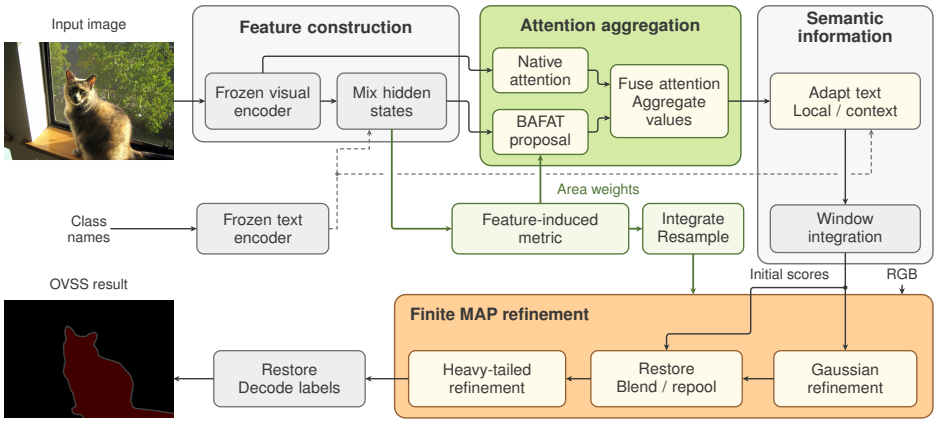

Figure 1: **RIMAR overview.** Mixed frozen states supply the final values and feature geometry; BAFAT and native attention determine their aggregation weights. Semantic information is integrated across configured windows, then refined using retained geometry and a separate RGB cue. Initial scores also enter the restored Gaussian blend before robust refinement. Both encoders remain frozen; dashed connections carry text prototypes.

3.1 SHARED FEATURE GEOMETRY

A uniform image grid does not reflect how rapidly visual features change across it. RIMAR uses this variation to define both patch contributions to attention and directional coupling in refinement. Let F be the channel-standardized patch field constructed below, with d channels and N locations, and let $J_i = [D_x F_i, D_y F_i]$ contain its spatial differences. The normalized metric and patch areas are

$$\widehat{g}_i = d^{-1} J_i^\top J_i, \qquad g_i = \widehat{g}_i / c, \qquad \omega_i = \frac{\sqrt{\det g_i}}{\sum_j \sqrt{\det g_j}}, \tag{1}$$

where c sets geometric-mean area density to one; Appendix A.2 specifies its calculation and positive stabilization. Area weights enter affine fitting and attention aggregation. For refinement, the retained metric instead supplies the local squared length $d^g_{ij} = \Delta u_{ij}^\top \frac{g_i^{\mathrm{ref}} + g_j^{\mathrm{ref}}}{2} \Delta u_{ij}$, which distinguishes edge directions without computing global geodesics. Figure 2 illustrates these complementary uses. Per-crop metric coefficients are averaged over covering windows, restored to the input raster, and pooled to the refinement grid. They are not recomputed from refined scores: sharing refers to the retained field across resolutions, not identical arrays or exact coordinate transport by interpolation.

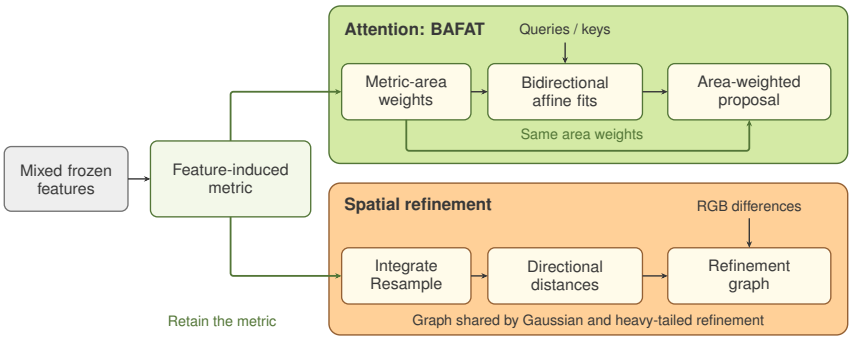

Figure 2: **One feature metric, two roles.** Normalized metric-area weights enter both bidirectional affine fitting and the BAFAT proposal, which subsequently joins native attention (Fig. 1). The retained metric coefficients are integrated and resampled to supply directional graph distances alongside RGB. The graph guides both finite refinement stages; geometry is not re-estimated from the refined scores.

3.2 Metric-Aware Attention and Feature Construction

Constructing the feature field. Intermediate states can retain spatial structure that is weaker in the last state, but uniform mixing can dilute useful features. RIMAR weights layers by attention concentration and agreement, using $\mathcal{I}_K(p) = \log K - H(p)$ and $\mathcal{B}(p,q) = \sum_j \sqrt{p_j q_j}$ for distributions over K tokens. These measure discrimination and overlapping support, not calibrated correctness. Comparing each native proposal with the last stored proposal gives

$$e_\ell = \mathrm{mean}_{\mathrm{heads,rows}} \left[\mathcal{I}_K(A^{(\ell)}) \mathcal{B}(A^{(\ell)}, A^{(M)})^2 \right], \qquad p = \mathrm{softmax}_\ell(e). \tag{2}$$

An interpolated contiguous support localizes p around its mean layer position, yielding $\widetilde{p}$. Class concentration from the late state and its native attention output controls the mixture

$$\rho = (1 - \beta_{\mathrm{cls}})p + \beta_{\mathrm{cls}}\widetilde{p}. \tag{3}$$

Appendix A.1 defines the support and semantic gate; intermediate layers receive no separate semantic scores. With normalized entropy $\bar{H}(\rho) = H(\rho)/\log M$, patches use

$$\bar{S}_i = (1 - \bar{H}(\rho)) S_i^{(M)} + \bar{H}(\rho) \sum_\ell \rho_\ell S_i^{(\ell)}, \qquad \bar{S}_0 = S_0^{(M)}.$$

Concentrated weights favor the last-state anchor, whereas diffuse weights admit more of the layer barycenter; the class token stays unchanged. This is an empirical design rule. Frozen Q/K/V projections act on $Y = \mathrm{LN}_1(\bar{S})$, while separate patch-channel standardization produces F for Eq. (1). This state mixture determines the features, values, and retained geometry; attention consensus then determines how those values are aggregated. The native attention proposals do not each contribute a separate value tensor.

Aligning and combining attention. BAFAT allows translation, scaling, and shear between query and key representations without training. For each head, it fits affine maps $T \in \mathcal{A}$ in both directions using metric-area weights:

$$T^\star_{\rightarrow} = \arg\min_{T \in \mathcal{A}} \sum_i \omega_i \|T(k_i) - q_i\|^2, \qquad T^\star_{\leftarrow} = \arg\min_{T \in \mathcal{A}} \sum_i \omega_i \|T(q_i) - k_i\|^2. \tag{4}$$

The aligned logits are separately softmaxed, added, area-reweighted, and normalized to form A^g; class-token logits remain native. The fits use Euclidean feature residuals, with closed-form solutions in Appendix A.3.

To retain the frozen model's spatial structure, RIMAR combines native proposals with A^g rather than replacing them. For each head and query row,

$$\begin{aligned} w_\ell &= M \rho_\ell \mathcal{I}_K(A^{(\ell)}) \mathcal{B}(A^{(\ell)}, A^g), \qquad A^{\mathrm{lay}} \propto \exp\left(\sum_\ell \frac{w_\ell}{\sum_j w_j} \log A^{(\ell)} \right), \\ \kappa_{\mathrm{lay}} &= M^{-1} \sum_\ell w_\ell, \quad \kappa_g = \mathcal{I}_K(A^g), \quad A^{\mathrm{final}} \propto \exp\left(\frac{\kappa_{\mathrm{lay}} \log A^{\mathrm{lay}} + \kappa_g \log A^g}{\kappa_{\mathrm{lay}} + \kappa_g} \right). \end{aligned} \tag{5}$$

Normalization is over keys. Concentration discounts diffuse rows, overlap discounts conflicting support, and log-probability pooling favors shared support. The resulting A^{final} aggregates the values projected from Y. This pooling of distributions is distinct from the spatial metric. Figure 3 separates feature construction, attention consensus, and the subsequent integration of semantic predictions across windows.

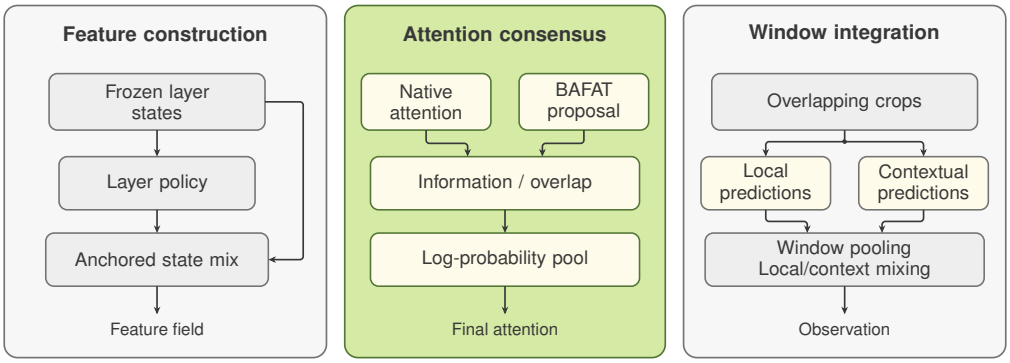

Figure 3: **Three aggregation decisions in Stage A.** RIMAR mixes frozen layer states with a final-state anchor, combines native attention with the metric-area BAFAT proposal using information, overlap, and log-probability pooling, and integrates local and contextual predictions across overlapping crops to form the observation.

3.3 Semantic Observations and Metric-Conditioned Refinement

From features to observations. Fixed text directions may fit some views better than others. RIMAR temporarily adapts each prototype using metric-area-weighted image-class covariance, projected onto the prototype's tangent plane and renormalized. The correction is reset for every view; encoder parameters remain frozen. Adapted patch similarities give local alias probabilities, while the class token and area-pooled class probabilities supply contextual support. Keeping both branches balances local detail against scene context. The contextual branch can support ambiguous local predictions, while retaining the local branch avoids replacing spatially varying scores by a single scene-wide distribution.

Overlapping crops can disagree, especially near their borders. Within each branch, RIMAR interpolates arithmetic and log-probability pooling according to crop agreement, then balances local and contextual predictions using image-level agreement and pixel-level concentration. Appendix A.4 specifies the distinct score conventions and gates. Full-image inference bypasses this integration; a one-crop sliding call still uses it. For the integrated probabilities P, A aliases, and a fixed score scale s_t, the observation is

$$X_a^0(u) = s_t^{-1}\left[\log P_a(u) - A^{-1}\sum_b \log P_b(u)\right]. \tag{6}$$

Centering retains relative alias preference and removes the class-independent offset.

Finite refinement. Figure 4 distinguishes the Gaussian correction, image-scale handoff, and robust updates.

RIMAR first restores X^0 to the input raster, retaining this notation, then pools the observation, RGB image, and metric to the posterior grid. Writing the observation there as X, it constructs

$$\alpha_i = \|\operatorname{softmax}(X_i) - U_A\|_2^2, \qquad D = \operatorname{diag}(\alpha), \qquad w_{ij} = \exp(-d_{ij}^I/s_I - d_{ij}^g/s_g).$$

Positive stabilization is specified in Appendix A.5. Here, U_A is the uniform alias distribution. The precision D anchors concentrated observations and stays fixed across both stages. Squared RGB differences d^I and feature lengths d^g independently weaken graph edges, with scales s_I, s_g given by their cardinal-edge medians.

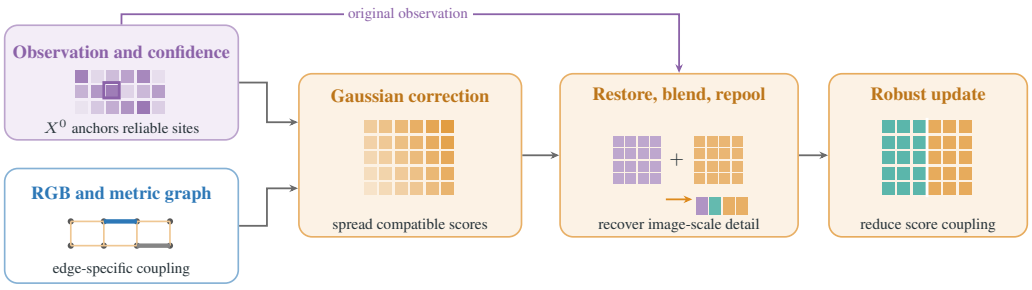

Figure 4: **Distinct roles in finite spatial refinement.** The observation and graph jointly determine the Gaussian correction. RIMAR then restores and blends that correction with the original observation before robust updates weaken coupling across large score differences.

The Gaussian stage applies a finite solve of $(D + L_w)Z_G = DX$, where L_w is the graph Laplacian. It reduces spatial noise before the robust stage weakens coupling across large score differences. Crucially, the Gaussian output is restored by $\mathcal{U}$, blended with the original observation, and repooled by $\mathcal{P}$:

$$B_H = \mathcal{P}[(1-\beta)X^0 + \beta\mathcal{U}\widetilde{Z}_G], \qquad w_{ij}^{(m)} = \frac{2w_{ij}}{1 + \|Z_i^{(m)} - Z_j^{(m)}\|^2/\tau^2}.$$

Starting from B_H, fixed robust rounds at scale τ update the weights and solve against DB_H with unchanged D. These finite updates follow Gaussian and heavy-tailed MAP objectives, not an assertion of exact global minimization. After restoration, scaled alias softmax, maximum-probability class reduction, and the dataset background rule yield pixel labels.

Algorithm 1 RIMAR inference with frozen encoders.

Require: Image I, text prototypes $\{t_a\}$ encoded once, alias map π, background threshold θ

Stage A: adaptive aggregation

for each full-image view or crop k of resized I **do**
  From frozen states and native attention, compute ρ_k and the anchored mixture $\bar{S}_k$.
  Project Q/K/V from normalized $\bar{S}_k$; derive $(g_k^{\mathrm{att}}, \omega_k)$ from standardized patches.
  Form BAFAT, fuse with native attention, and aggregate the projected values.
  Adapt temporary prototypes; form local and contextual alias probabilities.
end for
Integrate scores and metrics in sliding mode; bypass integration in full-image mode.
Form X^0 using Eq. (6); restore scores, RGB, and metric to the input raster.

Stage B: finite spatial refinement

Pool to the posterior grid; obtain observation X, fidelity D, and RGB/metric graph w.
$\widetilde{Z}_G \leftarrow \mathrm{PCG}(D + L_w, DX; X)$.
$B_H \leftarrow \mathcal{P}[(1-\beta)X^0 + \beta\mathcal{U}\widetilde{Z}_G]$.
Rebuild the graph; apply finite robust rounds from B_H with the original D, yielding Z.
return labels from $\mathrm{softmax}_a(s_t\mathcal{U}Z)$ using classwise alias maxima and threshold θ.

3.4 Coordinate Consistency of the Reference Family

The shared geometry also permits a coordinate-consistency statement, separate from prediction accuracy. Appendix B defines a continuous BAFAT-MAP family whose feature metric, sampling measure, fidelity, and regularization fields are transported together.

Theorem 1 (Coordinate consistency of the reference family). *For fixed $\gamma \geq 0$ and blend $\beta \in [0, 1]$, under the regularity, positivity, and transport assumptions of Appendix B, the semantic score, unique Gaussian minimizer, blended observation, heavy-tailed global-minimizer set, and induced label set transform by pullback. A singleton label set satisfies $\widehat{Y}' = \widehat{Y} \circ \phi$ almost everywhere.*

The result excludes adaptive aggregation, independent feature extraction, raster resampling, finite solves, and decoding; it does not imply camera invariance of the complete predictor.

4 EXPERIMENTS

4.1 SETUP AND COMPLETE-METHOD RESULTS

We use frozen OpenAI CLIP ViT-B/16 and evaluate 13 settings: WildPASS (Yang et al., 2021), Matterport3D (Chang et al., 2017), and Stanford2D3D (Armeni et al., 2017); FDD (Jang et al., 2022) and WoodScape (Yogamani et al., 2019); and VOC21/20 (Everingham et al., 2010), Context60/59 (Mottaghi et al., 2014), COCO Object/Stuff (Lin et al., 2014; Caesar et al., 2018), Cityscapes (Cordts et al., 2016), and ADE20K (Zhou et al., 2017). We average Stanford2D3D's three area-group scores first, then give all 13 settings equal weight. mIoU averages present classes; pAcc averages valid pixels. Appendix C specifies resizing, windows, and label conventions. All experiments utilize a single NVIDIA H100 GPU with 96GB memory.

In Table 1, RIMAR achieves 41.4% mean mIoU and 70.6% pAcc, exceeding PEARL by 2.9 and 3.8 points as well as all considered baselines. It leads 11 settings; SFP is higher on Matterport3D and CLIPtrase on COCO Object. The full pAcc table and domain means appear in Appendix C.

Table 1: mIoU (%) for training-free OVSS across 13 benchmark settings. Avg. is the reported arithmetic mean computed before display rounding. All methods use their stated inference procedures under the common dataset settings.

Method	Venue	Panoramic			Fisheye		Perspective								Avg.
		Wild	M3D	S2D	FDD	Wood	V21	PC60	Obj.	V20	PC59	Stuff	City	ADE	
CLIP	ICML'21	11.5	5.1	10.1	7.3	3.8	18.6	7.8	6.5	49.1	11.2	7.2	6.7	3.2	11.4
MaskCLIP	ECCV'22	34.4	17.7	27.0	22.3	20.4	38.3	23.6	20.6	74.9	26.4	16.4	12.6	9.8	26.5
GEM	CVPR'24	45.7	22.7	34.7	24.5	22.3	46.2	32.7	33.3	77.7	32.6	15.7	34.0	17.8	33.9
CLIPtrase	ECCV'24	45.1	22.9	35.2	24.5	22.6	50.9	29.9	**43.6**	81.0	33.8	22.8	21.3	16.4	34.6
ClearCLIP	ECCV'24	44.4	22.6	35.2	23.4	21.6	51.8	32.6	33.0	80.9	35.9	23.9	30.0	16.7	34.8
SCLIP	ECCV'24	43.0	20.4	33.1	24.0	21.3	59.1	30.4	30.5	80.4	34.1	22.4	32.2	16.1	34.4
NACLIP	WACV'25	47.3	24.2	36.1	26.2	24.6	64.1	35.0	36.2	83.0	38.4	25.7	35.5	19.1	38.1
SFP	ICCV'25	45.9	**25.7**	34.8	24.8	24.0	56.8	32.3	32.1	83.4	36.0	24.0	34.1	18.1	36.3
G-CDAM	ICLR'25	41.5	24.1	35.3	25.4	21.3	53.7	30.0	33.4	78.0	36.4	24.7	34.2	17.9	35.1
S-GLA	CVPR'26	42.8	17.8	23.2	17.5	18.2	60.1	31.0	33.9	78.9	34.0	22.9	20.3	14.9	32.0
PEARL	CVPR'26	44.1	23.9	37.9	25.7	23.4	64.1	35.1	37.3	86.9	38.6	26.3	37.6	19.4	38.5
RIMAR	-	**52.6**	25.2	**38.6**	**28.7**	**26.7**	**68.6**	**37.8**	39.2	**88.8**	**41.6**	**27.1**	**43.0**	**20.4**	**41.4**

4.2 UPSTREAM OUTPUTS AND REFINEMENT

Table 2 compares RIMAR and PEARL through one centered-log alias interface, with PEARL's native propagation removed. Both benefit from the same RGB-only refiner; RIMAR leads by 2.715 mIoU points before refinement and 2.203 afterward. This compares upstream systems without equalizing their native score calibration or testing shared geometry; paired intervals are in Appendix C.

Table 2: Common RGB refinement: mean mIoU/pAcc (%) over 13 settings.

Source / refiner	mIoU	pAcc
RIMAR / off	39.23	68.00
RIMAR / on	**41.41**	**70.66**
PEARL / off	36.51	64.87
PEARL / on	39.20	67.72

Gaussian-only refinement provides a useful stopping point: 41.078 mIoU and 70.115 pAcc. The subsequent blend, repooling, and robust rounds add 0.316 mIoU points while increasing refiner-only time from 4.35 to 20.11 ms/image on identical cached upstream fields. This comparison includes the complete handoff, so it does not isolate robust updates alone.

4.3 GEOMETRIC CONSTRUCTION AND SPATIAL GUIDANCE

A joint simplification retains adaptive aggregation but uses a self-self proposal, uniform metric-area quantities, identity-limit metric gates, and RGB-only refinement. Full RIMAR gains 0.444 mIoU points (95% paired interval $[0.401, 0.486]$); this compares the joint construction, not an isolated MAP-distance effect or attention-refinement synergy.

Table 3: **Accuracy-cost summary.** mIoU is the full 13-setting mean (Table 1; simplification from its control). Latency is synchronized predictor time with inputs/text resident; memory is peak allocated/reserved CUDA memory including one model.

Predictor	mIoU	Latency (ms/image)	Peak memory (GiB)
PEARL	38.5	153.17	6.81 / 18.26
Joint simplification	41.0	101.13	7.16 / 25.53
RIMAR	**41.4**	108.66	7.16 / 25.59

Table 3 uses 32 fixed images from each of 15 physical splits and ten synchronized, balanced blocks after warm-up, with identical inputs, precision, and crop-batch policies. Full RIMAR costs 7.53 ms/image more than the simplification and is faster than PEARL under this hardware and measurement scope.

With BAFAT area weighting and upstream processing fixed, feature distance adds 2.667 mIoU points without RGB but changes it by -0.014 with RGB (Table 4). An RGB graph rescaled to the same total edge mass scores 41.423, versus 41.394 for combined guidance; uniform coupling at that mass loses 7.019 points. Thus, spatial allocation matters beyond smoothing strength, while the extra feature-distance term gives no aggregate gain when RGB is active. Appendix C reports the intervals and interaction contrast.

Table 4: Spatial guidance: RGB (r) and feature distance (g), with BAFAT area weighting on. Values are 13-setting means (%).

r	g	mIoU	pAcc
0	0	34.302	64.949
1	0	**41.408**	**70.660**
0	1	36.969	67.338
1	1	41.394	70.643

4.4 Feature Construction and Layer Policy

Table 5 compares adaptive aggregation with one fixed counterpart using the same states, native proposals, and crops: uniform state/attention pooling and arithmetic window integration, with the full replacement defined in Appendix C. The adaptive construction gains 4.377 mIoU points over fixing both rules; the conditional layer and window effects are $+4.199$ and $+0.290$. This is a comparison with one declared fixed rule, not all possible fixed policies.

Table 5: Layer (L) and window (W) fusion: adaptive (1) or fixed (0). Values are 13-setting means (%) before and after refinement.

		Pre		Final	
L	W	mIoU	pAcc	mIoU	pAcc
1	1	39.23	68.00	**41.39**	**70.64**
0	1	36.20	65.78	37.19	67.73
1	0	39.06	67.72	41.10	70.30
0	0	36.08	65.44	37.02	67.34

A narrower policy-transfer study keeps row-local consensus and later adaptation active. A population policy calibrated on additional unlabeled images recovers most of the result; native per-view policies improve only slightly over it and over shuffling (Appendix C). The broader 4.377-point gain therefore cannot be attributed solely to input-specific layer selection. Close alignment substitutions are also near-tied: affine freedom has no established material advantage over orthogonal or self-self proposals.

The broader component-removal study appears in Appendix C.5.

4.5 Controlled View Changes and Sensitivity

Figure 5 summarizes paired views independently encoded from 772 Matterport3D and 373 Stanford2D3D panoramas. The complete predictor has higher mean accuracy and agreement than PEARL and NACLIP, with residual sensitivity to projection and sampling. Because component controls indicate that feature geometry alone does not account for an isolated consistency gain, we attribute these robustness improvements to the integrated inference framework rather than feature geometry in isolation.

Backbone ablation. Table 6 fixes input geometry while changing the complete CLIP checkpoint, jointly varying patch size, capacity, and checkpoint history. RIMAR leads both references on average with ViT-B/16 and ViT-L/14, but trails them with ViT-B/32.

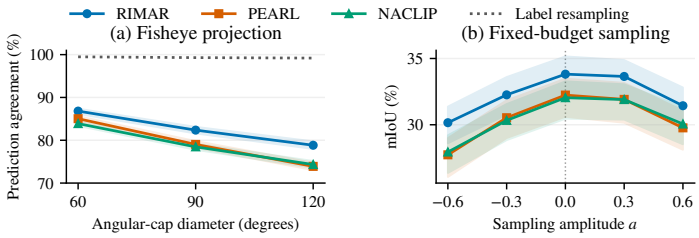

Figure 5: **Controlled view changes.** Left: agreement with each method's own same-cap perspective prediction after fisheye reprojection. Right: mIoU under fixed-field nonuniform sampling. Curves are equal-dataset means; shading gives pointwise 95% intervals from 1,000 paired panorama-bootstrap replicates.

Table 6: Backbone sensitivity: mIoU/pAcc over full FDD12, WildPASS, and VOC21 splits under identical window settings.

Backbone	RIMAR	PEARL	NACLIP
ViT-B/32	40.38/63.51	42.03/76.25	42.80/75.70
ViT-B/16	**49.19/83.43**	44.19/79.40	45.31/80.06
ViT-L/14	**49.90/83.93**	42.45/78.28	44.99/80.61

Figure 6 illustrates selected qualitative improvements across the three image domains. In these examples, RIMAR recovers more continuous road and vehicle regions and more complete object silhouettes than PEARL and NACLIP. Appendix C gives further results and bootstrap details.

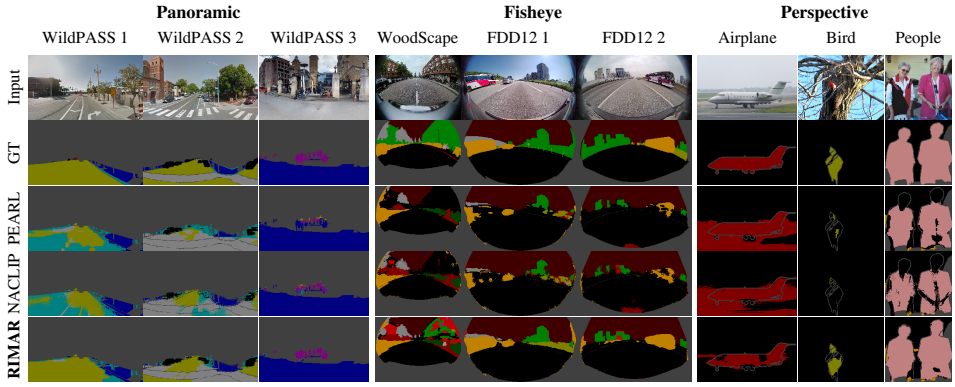

Figure 6: **Selected qualitative improvements.** Panoramic, fisheye, and perspective cases chosen from 768 images for large gains over the stronger reference; they do not represent improvement frequency. Gray is ignored.

5 Limitations and Conclusion

RIMAR links attention and refinement through feature-derived geometry: metric area weights its affine proposal, while metric lengths accompany RGB in refinement. The complete predictor constructs multi-layer features, integrates native and BAFAT attention, adapts to the input, merges window fields, and applies finite Gaussian and robust updates, reaching 41.4% mIoU and 70.6% pAcc across 13 settings. Its continuous reference family is coordinate-consistent under stated transport assumptions. Controls show a modest loss for the joint simplification, near-tied alignment choices, and no aggregate feature-distance gain with RGB, although feature distance helps without it. The updates are approximate, and cross-view gains reflect the integrated framework and complementary non-geometric aggregation, rather than geometry in isolation.

6 AI USE STATEMENT

In this work, we used generative AI tools to assist with mathematical formulation and proofs, hypothesis refinement, experimental design and execution, implementation, qualitative analysis, and interpretation of results. We have not used generative AI to create benchmark photographs or annotations; language translation and thematic analysis of interview or survey data are not applicable. Additionally, AI tools supported writing and editing, literature discovery, and figure preparation. All retained AI-assisted material was reviewed and verified by checking original sources and mathematical arguments, testing code, and comparing reported results with experimental records. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

7 REPRODUCIBILITY STATEMENT

The supplementary material provides anonymous RIMAR source code, benchmark configurations, and evaluation scripts. Please follow the instructions provided in the archive to reproduce the results from Table 1.

Section 3 describes the method, Appendix A supplies additional mathematical definitions, and Appendix C summarizes the evaluation protocol. Appendix B states the assumptions and provides the complete proof of the coordinate-consistency result.

REFERENCES

Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. Joint 2D-3D-semantic data for indoor scene understanding. *CoRR*, abs/1702.01105, 2017.

Yaniv Benny and Lior Wolf. SphereUFormer: A U-shaped transformer for spherical 360 perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 940-950, 2025.

Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3828-3837, 2024.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Oscar Carlsson, Jan E. Gerken, Hampus Linander, Heiner Spieß, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. HEAL-SWIN: A vision transformer on the sphere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6067-6077, 2024.

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

Qi Chen, Lingxiao Yang, Yun Chen, Nailong Zhao, Jianhuang Lai, Jie Shao, and Xiaohua Xie. Training-free class purification for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23124-23134, 2025.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4113-4123, 2024.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303-338, 2010.

Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 5061-5071, 2025.

Qiming Huang, Hao Ai, and Jianbo Jiao. Structure-aware feature rectification with region adjacency graphs for training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4118-4127, 2026.

Sujin Jang, Joohan Na, and Dokwan Oh. DaDA: Distortion-aware domain adaptation for unsupervised semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18681-18693, 2022.

Shuo Jin, Siyue Yu, Bingfeng Zhang, Mingjie Sun, Yi Dong, and Jimin Xiao. Feature purification matters: Suppressing outlier propagation for training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20291-20300, 2025.

Dong Un Kang, Hayeon Kim, and Se Young Chun. Class distribution-induced attention map for open-vocabulary semantic segmentations. In *International Conference on Learning Representations*, 2025.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ClearCLIP: Decomposing CLIP representations for dense vision-language inference. In *European Conference on Computer Vision*, 2024a.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ProxyCLIP: Proxy attention improves CLIP for open-vocabulary segmentation. In *European Conference on Computer Vision*, pp. 70-88, 2024b.

ByeongCheol Lee, Hyun Seok Seong, Sangeek Hyun, Gilhan Park, WonJun Moon, and Jae-Pil Heo. Looking beyond the window: Global-local aligned CLIP for training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27686-27696, 2026.

John M. Lee. *Introduction to Riemannian Manifolds*. Springer, 2nd edition, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

Kunal Mahatha, Jose Dolz, and Christian Desrosiers. NERVE: Neighbourhood & entropy-guided random-walk for training free open-vocabulary segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3243-3253, 2026.

Seungjae Moon, Seunghyun Oh, and Youngmin Ro. OV-Stitcher: A global context-aware framework for training-free open vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 7357-7367, 2026.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Gensheng Pei, Xiruo Jiang, Xinhao Cai, Tao Chen, Yazhou Yao, and Byeungwoo Jeon. PEARL: Geometry aligns semantics for training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17927-17937, 2026.

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 *International Conference on Machine Learning*, pp. 8748-8763, 2021.

Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of CLIP for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*, 2024.

Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. CLIP as RNN: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13171-13182, 2024.

Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, 2024.

Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. DeCLIP: Decoupled learning for open-vocabulary dense perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Kailun Yang, Jiaming Zhang, Simon Reiss, Xinxin Hu, and Rainer Stiefelhagen. Capturing omni-range context for omnidirectional segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1376-1386, 2021.

Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. ResCLIP: Residual attention for training-free dense vision-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29968-29978, 2025.

Senthil Yogamani, Ciaran Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricar, Stefan Milz, Martin Simon, Karl Amende, Christian Witt, Hazem Rashed, Sumanth Chennupati, Sanjaya Nayak, Saquib Mansoor, Xavier Perrotton, and Patrick Perez. WoodScape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9308-9318, 2019.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *European Conference on Computer Vision*, 2022.

A Additional Mathematical Details

This appendix completes the adaptive rules in Secs. 3.2-3.3 without repeating finite-precision execution details. Algorithm 1 gives the chronological computation, while Figs. 3 and 4 illustrate its aggregation and refinement steps.

Write $[x]_\varepsilon = \max(x, \varepsilon)$, $\mathrm{mass}(v) = v/[\sum_j v_j]_\varepsilon$, and let $\mathcal{C}$ sum alias probabilities into their assigned semantic classes.

A.1 Layer Distribution

Let layer positions be $r_\ell = \ell - 1$ and let $v_s = 4$. From the unrestricted distribution p in Eq. (2), set

$$\bar{r} = \sum_\ell r_\ell p_\ell, \qquad s = \mathrm{clip}\big(\bar{r} - \tfrac{1}{2}(v_s - 1), 0, M - v_s\big), \qquad k = \lfloor s \rfloor, \qquad f = s - k.$$

Let $m_\ell^- = \mathbf{1}_{\{k \le r_\ell < k + v_s\}}$ and $m_\ell^+ = \mathbf{1}_{\{k+1 \le r_\ell < k+1+v_s\}}$, clipping the second window at the final valid start. The interpolated support distribution is

$$\widetilde{p}_\ell = (1 - f)\frac{p_\ell m_\ell^-}{\sum_j p_j m_j^-} + f\frac{p_\ell m_\ell^+}{\sum_j p_j m_j^+}. \tag{7}$$

To form the semantic class mass, project a hidden sequence S through the frozen final normalization and visual projection, compute patch alias probabilities, sum aliases into classes, and take hard patch winners $\hat{c}_i(S)$. If $\omega_i(S)$ is the area distribution induced by the same unprojected state, define

$$\gamma_c(S) = \sum_i \omega_i(S)\mathbf{1}_{\{c = \hat{c}_i(S)\}}. \tag{8}$$

Use $u = \gamma(S^{(M)})$ and $v = \gamma(S^{\mathrm{att}})$, where S^{att} is the native final-block attention output on $S^{(M)}$ before residual and MLP. With $J_u = \mathcal{I}_C(u)$, $J_v = \mathcal{I}_C(v)$, and $a = J_u/(J_u + J_v)$ (or $1/2$ when both vanish), form

$$m^{\mathrm{mix}} = \mathrm{mass}\Big(\big[a\sqrt{u} + (1-a)\sqrt{v}\big]^2\Big), \qquad m^{\mathrm{prod}} = \mathrm{mass}(uv), \tag{9}$$

$$o = \mathcal{B}(u, v)^2, \qquad m = \mathrm{mass}\big(o m^{\mathrm{prod}} + (1 - o)m^{\mathrm{mix}}\big). \tag{10}$$

For support width v_s, the semantic-concentration gate is

$$b_{\mathrm{cls}}(x) = \frac{(v_s - 1)x}{1 + (v_s - 2)x}, \qquad \beta_{\mathrm{cls}} = b_{\mathrm{cls}}\left(\sum_c m_c^2\right).$$

This empirical monotone rule increases the contribution of the localized distribution as the class mass becomes more concentrated. Equations (7)-(10), this gate, and Eq. (3) fully define ρ. Only the late state and its native attention output enter m. The contiguous support favors a coherent region of the transformer rather than a discontinuous top-k choice. Interpolating adjacent supports avoids abrupt replacement as the mean position changes and can retain five layers; the final mixture also retains unrestricted mass. The support, semantic gate, and last-state anchor are empirical design conventions rather than consequences of the Riemannian metric.

A.2 Finite Feature Metric

For a requested hidden sequence, remove the class token, standardize each channel across its N patches, and reshape it to the patch grid. With replicated boundary values, use centered differences

$$D_x F(y, x) = \tfrac{1}{2}[F(y, x+1) - F(y, x-1)], \qquad D_y F(y, x) = \tfrac{1}{2}[F(y+1, x) - F(y-1, x)].$$

Thus, boundary derivatives are half one-sided differences and coordinates are measured in patches. The unnormalized coefficients are

$$\widehat{g}_{xx} = d^{-1}\sum_c (D_x F_c)^2, \quad \widehat{g}_{yy} = d^{-1}\sum_c (D_y F_c)^2, \quad \widehat{g}_{xy} = d^{-1}\sum_c D_x F_c D_y F_c.$$

Positive stabilization is applied to the diagonal coefficients before the determinant normalization. The scale in Eq. (1) is

$$c = \exp\left(\frac{1}{2N}\sum_i \log\det \widehat{g}_i\right),$$

so the normalized metric has geometric-mean area density one and area weights proportional to $\sqrt{\det g_i}$. These weights measure feature-induced area rather than physical scene area or camera solid angle. When an adaptive gate requests scalar metric information, RIMAR uses

$$\mathcal{I}_g(i) = 1 - \exp[-D_g(i)], \qquad D_g(i) = \tfrac{1}{2}[\operatorname{tr} g_i - \log\det g_i - 2]_+.$$

This is the Gaussian divergence of the local metric from the native identity metric, compressed to $[0, 1)$. Stabilizing floors keep the finite metric positive; they do not introduce camera calibration or physical coordinates.

To pass the retained metric to refinement, let $\mathcal{W}_u$ contain the crops covering pixel u, $\mathcal{U}_k$ map crop coefficients into their image locations by bilinear interpolation, $\mathcal{R}$ restore the original image size, and $\mathcal{P}$ pool to the posterior grid. Then

$$g_k^{\mathrm{att}} = \mathrm{Metric}(F_k), \qquad G(u) = \frac{1}{|\mathcal{W}_u|}\sum_{k\in\mathcal{W}_u}(\mathcal{U}_k g_k^{\mathrm{att}})(u), \qquad g^{\mathrm{ref}} = \mathcal{P}\mathcal{R}(G).$$

Full-image inference retains a single metric field; sliding inference integrates every covering view. The coefficients are integrated and resampled rather than recomputed from averaged features or refined scores. Auxiliary metric evaluations used by preliminary semantic and layer gates can differ from this retained final-attention field. Coefficient interpolation in the finite predictor is not asserted to be exact tensor transport under arbitrary coordinate changes.

A.3 Weighted Affine Solution and Proposal

The affine fits use Euclidean query/key residuals weighted by patch area; the spatial 2×2 metric is not a high-dimensional Mahalanobis norm. For paired row vectors (k_i, q_i) with positive normalized weights ω_i, let $\mu_k = \sum_i \omega_i k_i$, $\mu_q = \sum_i \omega_i q_i$, and $C_{ab} = \sum_i \omega_i (a_i - \mu_a)^\top (b_i - \mu_b)$. A weighted least-squares solution to the forward fit in Eq. (4) is

$$T_{\rightarrow}(k) = (k - \mu_k)C_{kk}^{\dagger}C_{kq} + \mu_q, \qquad T_{\leftarrow}(q) = (q - \mu_q)C_{qq}^{\dagger}C_{qk} + \mu_k.$$

The implementation first attempts the corresponding covariance solve and uses a symmetric pseudoinverse only if that solve fails. For head width d_h, the patch logits are $\ell_{ij}^{\rightarrow} = q_i T_{\rightarrow}(k_j)^\top/\sqrt{d_h}$ and $\ell_{ij}^{\leftarrow} = T_{\leftarrow}(q_i)k_j^\top/\sqrt{d_h}$. Replacing the class row and column by native logits, the proposal used in Eq. (5) is

$$A_{ij}^{g} \propto \left[\mathrm{softmax}_j(\ell_{ij}^{\rightarrow}) + \mathrm{softmax}_j(\ell_{ij}^{\leftarrow})\right]\begin{cases}1, & j = 0,\\ N\omega_j, & j \geq 1.\end{cases}$$

A.4 Semantic and Window Weights

Temporary prototype adaptation. For normalized output patches z_i and text prototypes t_a, initial alias probabilities are $R_{ia} \propto \exp(s_t z_i^\top t_a)$, using the fixed semantic scale s_t from Eq. (6). Sum aliases to classes as $R^{\mathrm{cls}} = \mathcal{C}R$ and let μ_i be normalized metric-area weights. RIMAR forms

$$m_c = \sum_i \mu_i R_{ic}^{\mathrm{cls}}, \qquad B_c = \sum_i \mu_i R_{ic}^{\mathrm{cls}} z_i - m_c \sum_i \mu_i z_i,$$

$$t_a' = \mathrm{unit}\Big[t_a + B_{\pi(a)} - (B_{\pi(a)}^\top t_a)t_a\Big].$$

The covariance term moves a prototype toward image features associated with its class. Its tangent-plane projection and renormalization preserve the unit-vector representation convention. The adapted prototypes are discarded after the current image or crop, and both encoders remain frozen. Similarities to t_a' produce the local probability field P^{loc}. The contextual branch adds a crop-level class prior to help resolve locally ambiguous categories, while retaining the local branch avoids forcing that prior onto every pixel.

Contextual probabilities. For one crop, class-token probabilities r^0, local class field $L = \mathcal{C}P^{\mathrm{loc}}$, and normalized area weights ν_u give

$$\ell_c = \sum_u \nu_u L_{cu}, \qquad q = \mathrm{mass}(\mathcal{C}r^0 + \ell), \qquad \varrho_a \propto q_{\pi(a)}.$$

With class entropy H, define global/local information $e_g = \log C - H(q)$ and $e_u = \log C - H(L_{\cdot u})$, global agreement $a_g = \mathcal{B}(q, \ell)^2$, concentration $\chi = \sum_c \ell_c^2$, and

$$\widehat{e}_g = e_g a_g^{1/\chi}, \qquad \gamma_u = \frac{\widehat{e}_g}{\widehat{e}_g + e_u}, \qquad P^{\mathrm{ctx}}_{au} = (1-\gamma_u)P^{\mathrm{loc}}_{au} + \gamma_u \varrho_a.$$

All nonnegative quantities use fixed numerical floors; they do not introduce learned parameters.

Window integration. Overlapping crops can disagree near their boundaries and differ in concentration. For each branch $b \in \{\mathrm{loc}, \mathrm{ctx}\}$, let $\bar{P}^b_u$ be the arithmetic mean over crops covering u. The branch and local/context mixtures are

$$Q^b_u = s^b_u Q^{b,\mathrm{geo}}_u + (1 - s^b_u)\bar{P}^b_u, \qquad P_u = (1-\Gamma_u)Q^{\mathrm{loc}}_u + \Gamma_u Q^{\mathrm{ctx}}_u. \tag{11}$$

Here, s^b_u favors shared-support pooling when crops agree and arithmetic averaging when they conflict. Let $c_{ka}(u)$ be crop k's bilinearly interpolated cosine score against an adapted prototype, and write

$$L_{ka}(u) = s_t^{-1}\left[\log P^{\mathrm{ctx}}_{ka}(u) - A^{-1}\sum_b \log P^{\mathrm{ctx}}_{kb}(u)\right].$$

The geometric candidates use different score conventions:

$$Q^{\mathrm{loc,geo}}_u = \mathrm{softmax}_a\left(\frac{s_t}{|\mathcal{W}_u|}\sum_{k\in\mathcal{W}_u} c_{ka}(u)\right), \qquad Q^{\mathrm{ctx,geo}}_u = \mathrm{softmax}_a\left(\frac{s_t}{|\mathcal{W}_u|}\sum_{k\in\mathcal{W}_u} L_{ka}(u)\right).$$

Thus, the local branch averages raw cosine scores, while the contextual branch averages centered-log scores before softmax. Full-image inference bypasses these window operations.

For overlapping crops, let M^b_{cu} and V^b_{cu} be the mean and variance of class probability at u, and n_u the number of covering crops. The agreement in Eq. (11) is

$$d^b_u = \begin{cases} \sum_c (M^b_{cu})^2 - \dfrac{\sum_c V^b_{cu}}{n_u - 1}, & n_u > 1, \\ \sum_c (M^b_{cu})^2, & n_u = 1, \end{cases} \qquad s^b_u = \left[\frac{d^b_u - 1/C}{1 - 1/C}\right]_0^1.$$

Let $K^b = \mathcal{C}Q^b$. The image-level branch comparison uses the area-weighted average

$$F = \sum_u \nu_u \{H(K^{\mathrm{loc}}_{\cdot u}) - H(K^{\mathrm{ctx}}_{\cdot u}) - D_{\mathrm{KL}}(K^{\mathrm{ctx}}_{\cdot u} \| K^{\mathrm{loc}}_{\cdot u})\},$$

$$\zeta = \mathrm{sigmoid}(N_{\mathrm{eff}} F), \qquad \theta_u = \frac{c^{\mathrm{ctx}}_u}{c^{\mathrm{loc}}_u + c^{\mathrm{ctx}}_u}, \qquad c^b_u = \left[\frac{\sum_c (K^b_{cu})^2 - 1/C}{1 - 1/C}\right]_0^1.$$

Here, N_{eff} is the inverse concentration of coarse area weights. For the averaged metric $\bar{G}_u$, define

$$j_u = \tfrac{1}{2}[\mathrm{tr}\,\bar{G}_u - \log\det \bar{G}_u - 2]_+, \qquad \eta_g = \sum_u \nu_u (1 - e^{-j_u}).$$

Then, $\Gamma_u = (1-\eta_g)\zeta + \eta_g \theta_u$ in Eq. (11). Thus, ζ is image-level, θ_u is pixel-level, and η_g selects their relative influence. A one-crop sliding call has $n_u = 1$ but still evaluates these equations.

A.5 Finite Refiner Algebra and Decoding

Figure 4 distinguishes the Gaussian correction, the image-scale handoff, and the robust updates.

The integrated score field, RGB image, and retained metric are restored to the original image raster before posterior-grid pooling. For the pooled observation $B_G = \mathcal{P}X^0$, the original fidelity is

$\alpha_i = \| \operatorname{softmax}_a(B_{G,i}) - U_A \|_2^2$, with a fixed positive floor and uniform alias distribution U_A. This rule uses the score field directly, without reapplying s_t, and remains fixed through both stages.

The four-neighbor-direction graph contains horizontal, vertical, down-right, and down-left undirected edges once. On the pooled grid,

$$d_{ij}^I = \tfrac{1}{3}\|I_i - I_j\|^2, \qquad d_{ij}^g = \Delta u_{ij}^\top \frac{g_i + g_j}{2} \Delta u_{ij}, \qquad w_{ij} = \exp(-d_{ij}^I / s_I - d_{ij}^g / s_g).$$

The cardinal-edge medians s_I, s_g are also used for diagonal edges. Let $D = \operatorname{diag}(\alpha)$ and L_w be the symmetric graph Laplacian. For an observation B, the objectives are

$$E_\rho^\gamma(Z; B) = \tfrac{1}{2} \sum_i \alpha_i \|Z_i - B_i\|^2 + \sum_{(i,j)} w_{ij} \rho(\|Z_i - Z_j\|^2) + \tfrac{\gamma}{2}\|Z\|_F^2, \tag{12}$$

with $\rho_G(s) = s/2$ and $\rho_H(s) = \tau^2 \log(1 + s/\tau^2)$. The Gaussian penalty first spreads compatible scores across the graph. The heavy-tailed penalty reduces coupling across large current score differences, which need not coincide with true semantic boundaries. Its quadratic surrogate uses

$$w_{ij}^{(m)} = \frac{2 w_{ij}}{1 + \|Z_i^{(m)} - Z_j^{(m)}\|^2 / \tau^2}.$$

The factor two follows from differentiating the heavy-tailed penalty with each undirected edge counted once. The exact Gaussian member of Eq. (12) would satisfy

$$(D + L_w + \gamma I) Z_G = D B_G.$$

At the evaluated $\gamma = 0$, the fidelity floor alone makes this matrix positive definite:

$$v^\top (D + L_w) v = \sum_i \alpha_i v_i^2 + \sum_{\{i,j\}} w_{ij} (v_i - v_j)^2 \geq 10^{-6} \|v\|^2.$$

The predictor returns 12 PCG iterations rather than the exact solution.

Write $\mathcal{P}$ for adaptive pooling to 112×112 and $\mathcal{U}$ for bilinear restoration. Its actual stage inputs and updates are

$$\begin{aligned}
&B_G = \mathcal{P} X^0, \qquad \widetilde{Z}_G = \mathrm{PCG}_{12}(D + L_w, D B_G; B_G), \\
&B_H = \mathcal{P}[(1-\beta) X^0 + \beta \mathcal{U} \widetilde{Z}_G], \qquad \beta = \tfrac{1}{2}, \\
&Z^{(0)} = B_H, \qquad Z^{(m+1)} = \mathrm{PCG}_{12}(D + L_{w^{(m)}}, D B_H; Z^{(m)}), \quad m = 0, \ldots, 4.
\end{aligned}$$

The heavy stage recomputes its base graph after restoration and repooling but retains the original D. In general, $\mathcal{P}\mathcal{U} \neq I$, so B_H cannot be replaced by $(1-\beta) B_G + \beta \widetilde{Z}_G$.

After restoring $Z^{(5)}$, RIMAR computes alias probabilities $\widehat{P}_a = \operatorname{softmax}_a(s_t \mathcal{U} Z^{(5)})$, reduces aliases by $R_c = \max_{a:\pi(a)=c} \widehat{P}_a$, and predicts the maximizing class unless $\max_c R_c$ is below the dataset background threshold. Intermediate class masses sum aliases; final decoding instead takes their maximum. The fixed finite iterates define the predictor and are not claimed to be exact global MAP minimizers.

B Reference Analysis

This appendix defines the continuous object in Theorem 1. It deliberately omits the adaptive layer, semantic, and window rules, sampled resampling, finite linear solves, and decoding.

B.1 Continuous Construction and Assumptions

Let $\mathcal{M}$ be a compact connected two-dimensional manifold with normalized sampling measure $\mathrm{d}\nu_s$. Let the standardized feature field $F : \mathcal{M} \to \mathbb{R}^d$ be continuously differentiable, have positive channel variances, and satisfy $\operatorname{rank} \mathrm{d}F = 2$. Then

$$g = d^{-1} \mathrm{d}F^\top \mathrm{d}F, \qquad \mathrm{d}\nu_g = \frac{\mathrm{d}\mu_g}{\int_{\mathcal{M}} \mathrm{d}\mu_g}$$

is a Riemannian metric with normalized volume. We assume that ν_s and ν_g are equivalent and that the density $r_s = \mathrm{d}\nu_s/\mathrm{d}\nu_g$ satisfies $0 < r_- \le r_s \le r_+ < \infty$. Frozen maps Q, K, V applied to an input feature field give q, k, v. Replacing sums in Eq. (4) by ν_g integrals defines $T^\star_\rightarrow, T^\star_\leftarrow$. With logits $\ell_\rightarrow(u,w)$ and $\ell_\leftarrow(u,w)$, for $\diamond \in \{\rightarrow, \leftarrow\}$ define the sampling-measure attention normalizers

$$Z_\diamond(u) = \int_{\mathcal{M}} e^{\ell_\diamond(u,w)}\, \mathrm{d}\nu_s(w).$$

The patch-only continuous BAFAT kernel is

$$a_g(u,w) = \frac{e^{\ell_\rightarrow(u,w)}/Z_\rightarrow(u) + e^{\ell_\leftarrow(u,w)}/Z_\leftarrow(u)}{\int_{\mathcal{M}} [e^{\ell_\rightarrow(u,z)}/Z_\rightarrow(u) + e^{\ell_\leftarrow(u,z)}/Z_\leftarrow(u)]\, \mathrm{d}\nu_g(z)}.$$

The visual field is $x(u) = \psi(\int a_g(u,w) v(w)\, \mathrm{d}\nu_g(w))$, where ψ is frozen. Fixed class embeddings $t_c \neq 0$ then give a cosine score field X. We assume $\inf_{u\in\mathcal{M}} \|x(u)\|_2 > 0$, so every cosine score is well defined.

Let α_X be a pointwise fidelity rule that commutes with composition and satisfies $0 < \alpha_- \le \alpha_X(u) \le \alpha_+ < \infty$ (the confidence rule with a positive floor is one example). For four continuously differentiable direction fields ξ_r, assume $0 < w_- \le w_r(u) \le w_+ < \infty$ and uniform tangent control: $\sum_r w_r(u) |\zeta(\xi_r(u))|^2 \ge c|\zeta|^2_{g^{-1}}$ for some $c > 0$. On $H^1(\mathcal{M}; \mathbb{R}^C)$ define

$$\mathcal{E}^\gamma_G[Z; X] = \tfrac{1}{2} \int \alpha_X \|Z - X\|^2\, \mathrm{d}\nu_s + \tfrac{1}{2} \sum_r \int w_r \|\mathrm{d}Z(\xi_r)\|^2\, \mathrm{d}\nu_s + \tfrac{\gamma}{2} \int \|Z\|^2\, \mathrm{d}\nu_s,$$

$$\begin{aligned}\mathcal{E}^\gamma_{H,X}[Z \mid W] = {} & \tfrac{1}{2} \int \alpha_X \|Z - W\|^2\, \mathrm{d}\nu_s \\ & + \sum_r \int w_r \tau^2 \log\!\left(1 + \frac{\|\mathrm{d}Z(\xi_r)\|^2}{\tau^2}\right) \mathrm{d}\nu_s + \tfrac{\gamma}{2} \int \|Z\|^2\, \mathrm{d}\nu_s. \end{aligned} \tag{13}$$

Positive fidelity makes the Gaussian objective coercive and strictly convex for every $\gamma \ge 0$, hence it has a unique minimizer $Z^\star_G$. For fixed $\beta \in [0,1]$, set $B_H = (1-\beta)X + \beta Z^\star_G$ and let $\mathcal{Z}_H(B_H)$ be the set of global minimizers of Eq. (13). The theorem makes no general heavy-stage existence or uniqueness claim: this set, and hence its induced label set, may be empty or non-singleton.

Lemma 1 (Gaussian well-posedness). *Under the assumptions above, $\mathcal{E}^\gamma_G[\cdot; X]$ has a unique global minimizer for every $\gamma \ge 0$.*

Proof. Boundedness of X on the compact domain and the positive fidelity bound $\alpha_X \ge \alpha_- > 0$ imply

$$\tfrac{1}{2} \int \alpha_X \|Z - X\|^2\, \mathrm{d}\nu_s \ge \tfrac{\alpha_-}{4} \|Z\|^2_{L^2} - C_X.$$

Uniform tangent control bounds the H^1 seminorm by the directional quadratic term, so the objective is coercive on H^1 even when $\gamma = 0$. Its quadratic terms are weakly lower semicontinuous; the direct method gives existence. The positive fidelity term is strictly convex, and the remaining terms are convex, giving uniqueness. □

For a fixed observation, the continuous heavy objective uses the same α_X but replaces the quadratic derivative penalty by the logarithmic one. That penalty is nonconvex in the derivative magnitude and does not give the preceding general coercivity argument. We therefore define

$$\mathcal{Y}_H(B_H) = \{\arg\max_{c=1,\dots,C} Z_c : \; Z \in \mathcal{Z}_H(B_H)\},$$

with fields and labels identified almost everywhere, without asserting that either set is nonempty or a singleton.

B.2 Transport and Proof

Let $\phi : \mathcal{M}' \to \mathcal{M}$ be a diffeomorphism and write $P_\phi f = f \circ \phi$. The primed construction transports

$$F' = P_\phi F, \quad \mathrm{d}\nu'_s = \phi^* \mathrm{d}\nu_s, \quad \alpha' = P_\phi \alpha, \quad w'_r = P_\phi w_{\sigma(r)}, \quad \mathrm{d}\phi\, \xi'_r = \varepsilon_r\, P_\phi \xi_{\sigma(r)}, \tag{14}$$

for a direction permutation σ and signs $\varepsilon_r \in \{-1, 1\}$. More generally, it suffices that the Gaussian and heavy regularization densities are natural under pullback. All frozen feature and class maps, γ, β, τ, and the fidelity convention remain fixed.

Proposition 1 (Metric and BAFAT naturality). *Under Eq.* (14)*, $g' = \phi^* g$, $\mathrm{d}\nu_{g'} = \phi^* \mathrm{d}\nu_g$, and the continuous semantic score satisfies $X' = P_\phi X$.*

Proof. The chain rule gives $\mathrm{d}F' = \mathrm{d}F \circ \mathrm{d}\phi$, hence the metric identity. Riemannian volume is natural under pullback, and normalization does not change because the total volumes agree by change of variables. Transport of ν_s also makes channel standardization commute with P_ϕ. Changing variables in every weighted affine moment therefore gives the same feature-coordinate affine maps. Changing variables once more in the two attention normalizers and value integral gives $x' = P_\phi x$; fixed cosine scoring then gives the score identity. □

For completeness, the regularization assumption can be written without choosing a particular direction permutation. For an $\mathbb{R}^C$-valued covector ζ, define densities

$$\Gamma_u(\zeta) = \tfrac{1}{2} \sum_r w_r(u) \|\zeta(\xi_r(u))\|^2,$$
$$\Psi_u(\zeta) = \sum_r w_r(u) \tau^2 \log\left(1 + \frac{\|\zeta(\xi_r(u))\|^2}{\tau^2}\right).$$

It is enough to assume

$$\Gamma'_{u'}(\zeta \circ \mathrm{d}\phi_{u'}) = \Gamma_{\phi(u')}(\zeta), \qquad \Psi'_{u'}(\zeta \circ \mathrm{d}\phi_{u'}) = \Psi_{\phi(u')}(\zeta). \tag{15}$$

Equation (14) is a concrete sufficient condition.

Proof of Theorem 1. Proposition 1 gives $X' = P_\phi X$. The fidelity rule commutes with composition, hence $\alpha'_{X'} = P_\phi \alpha_X$. For every $Z \in H^1(\mathcal{M}; \mathbb{R}^C)$, the weak chain rule, Eq. (15), and change of variables yield

$$\mathcal{E}'^{\gamma}_G[P_\phi Z; P_\phi X] = \mathcal{E}^\gamma_G[Z; X]. \tag{16}$$

Lemma 1 supplies the unique Gaussian minimizers. Since P_ϕ is a bijection on H^1, Eq. (16) implies $Z'^\star_G = P_\phi Z^\star_G$. The fixed scalar blend commutes with pullback: $B'_H = (1-\beta) P_\phi X + \beta P_\phi Z^\star_G = P_\phi B_H$.

The same weak-chain-rule and change-of-variables argument gives

$$\mathcal{E}'^{\gamma}_{H,X'}[P_\phi Z \mid P_\phi B_H] = \mathcal{E}^\gamma_{H,X}[Z \mid B_H].$$

Bijectivity therefore maps the complete global-minimizer set: $\mathcal{Z}'_H(B'_H) = P_\phi \mathcal{Z}_H(B_H)$. Pointwise class selection commutes with composition, so the induced label sets also pull back. If the label set is a singleton, its member obeys $\widehat{Y}' = \widehat{Y} \circ \phi$ almost everywhere. □

The evaluated predictor uses $(\gamma, \beta) = (0, 1/2)$ but inserts image-grid restoration and repooling between finite iterations. Sharing these scalar settings does not identify it with the reference minimizers. In particular, the proof does not cover independently rendered RGB inputs whose ViT features are extracted anew.

C Additional Results and Evaluation Notes

C.1 Benchmark Settings

Table 7: Benchmark settings. C/A denotes semantic classes/ordered aliases, s the target short side, and θ the background threshold. Every resize uses a 2048-pixel long-edge cap.

Setting	Split	Images	C/A	s	Inference	θ
WildPASS	validation	500	8/8	640	224/112	0.0
Matterport3D	validation	772	20/20	640	224/112	0.0
Stanford2D3D	three area groups	483/557/373	13/13	640	224/112	0.0
FDD	validation	974	12/12	640	224/112	0.0
WoodScape	validation	1000	17/17	640	224/112	0.0
VOC21	VOC2012 validation	1449	21/56	336	full image	0.1
Context60	VOC2010 validation	5104	60/60	336	224/112	0.1
COCO Object	COCO2017 validation	5000	81/113	336	224/112	0.1
VOC20	VOC2012 validation	1449	20/30	336	full image	0.0
Context59	VOC2010 validation	5104	59/59	336	224/112	0.0
COCO Stuff	COCO2017 validation	5000	171/171	336	224/112	0.0
Cityscapes	validation	500	19/19	560	224/112	0.0
ADE20K	ADE2016 validation	2000	150/150	336	224/112	0.0

The two VOC label spaces share the same 1,449 images, the two Context spaces share 5,104 images, and the two COCO spaces share 5,000 images. Stanford2D3D is scored in its three physical area groups and averaged equally into one logical setting. A crop/stride entry of 224/112 uses overlapping windows; full-image inference bypasses their integration. Raw ignore labels and each dataset's declared void pixels are excluded from both metrics.

C.2 Pixel Accuracy and Component Detail

Table 1 compares CLIP and MaskCLIP (Radford et al., 2021; Zhou et al., 2022), GEM, CLIPtrase, ClearCLIP, and SCLIP (Bousselham et al., 2024; Shao et al., 2024; Lan et al., 2024a; Wang et al., 2024), NACLIP, SFP, and G-CDAM (Hajimiri et al., 2025; Jin et al., 2025; Kang et al., 2025), S-GLA (Lee et al., 2026), and PEARL (Pei et al., 2026).

Table 8: Pixel accuracy (pAcc, %) across the same 13 settings as Table 1. Avg. is the reported arithmetic mean computed before display rounding.

Method	**Venue**	**Panoramic**			**Fisheye**		**Perspective**							**Avg.**
		Wild	**M3D**	**S2D**	**FDD**	**Wood**	**V21**	**PC60**	**Obj.**	**V20**	**PC59**	**Stuff**	**City**	**ADE**	
CLIP	ICML'21	25.6	13.0	33.0	36.5	18.7	53.7	23.0	62.9	65.8	26.5	13.0	29.5	7.8	31.5
MaskCLIP	ECCV'22	49.4	44.4	53.7	74.5	51.4	42.8	46.7	30.1	84.4	50.6	31.7	43.7	25.2	48.4
GEM	CVPR'24	71.8	51.9	63.6	76.3	55.7	36.9	52.6	49.2	86.0	60.3	40.5	71.2	43.5	58.4
CLIPtrase	ECCV'24	71.2	52.0	64.8	77.5	57.9	78.6	52.1	50.1	89.7	58.9	38.9	63.4	38.6	61.1
ClearCLIP	ECCV'24	71.1	52.0	65.9	75.4	55.8	82.4	50.8	73.7	88.8	59.3	39.7	69.3	40.2	63.4
SCLIP	ECCV'24	65.7	47.4	65.4	77.4	58.1	87.6	49.2	74.3	91.0	58.3	38.4	72.7	38.7	63.4
NACLIP	WACV'25	72.9	52.3	65.4	74.5	53.6	89.3	53.5	77.1	90.5	62.8	41.7	68.8	48.1	65.4
SFP	ICCV'25	71.4	57.3	60.5	75.6	55.0	85.3	53.6	71.3	91.9	60.7	41.3	71.7	43.8	64.6
G-CDAM	ICLR'25	67.9	52.9	60.1	77.2	56.1	82.8	43.5	70.2	87.1	60.8	42.1	73.0	45.4	63.0
S-GLA	CVPR'26	66.7	46.1	46.2	66.4	49.8	87.8	49.5	74.8	89.4	57.1	37.5	45.4	39.4	58.2
PEARL	CVPR'26	71.4	53.3	69.5	78.1	58.5	88.5	53.5	75.8	93.3	62.9	42.9	73.8	46.9	66.8
RIMAR	-	**78.2**	**57.4**	**69.7**	**81.0**	**62.8**	**91.2**	**57.7**	**80.3**	**94.4**	**66.9**	**45.6**	**81.0**	**52.3**	**70.6**

The headline panoramic, fisheye, and perspective mIoU means are 38.8, 27.7, and 45.8.

Table 9 gives one compact per-setting view of the two central mechanism studies. The adaptive-versus-fixed aggregation result is broad across settings, whereas the common-refiner comparison

is not uniform: PEARL is higher on Stanford2D3D after common refinement. The conclusions in Secs. 4.2 and 4.4 therefore concern fixed-setting means and paired differences, rather than universal per-dataset dominance.

Table 9: Per-setting mIoU for adaptive versus fixed aggregation and for RIMAR versus PEARL with a common RGB refiner. The fixed variant uses uniform layer and window rules; the common-refiner comparison removes PEARL's native propagation and applies the same downstream RGB refinement and decoding.

	Aggregation		Common RGB refiner	
Setting	Adaptive	Fixed	RIMAR	PEARL
WildPASS	52.60	49.11	51.87	45.19
Matterport3D	25.15	24.50	25.05	24.52
Stanford2D3D	38.63	33.81	38.53	39.28
FDD12	28.70	24.63	28.81	26.56
WoodScape17	26.68	23.24	26.56	24.22
VOC21	68.62	56.44	69.24	65.10
Context60	37.75	34.75	37.90	35.85
COCO Object	39.20	33.37	39.37	37.77
VOC20	88.75	82.49	88.76	87.09
Context59	41.59	38.36	41.74	39.44
COCO-Stuff	27.08	24.44	27.21	26.75
Cityscapes	42.94	38.19	42.70	38.04
ADE20K	20.44	17.89	20.55	19.84

Aggregate contrasts and finite refinement. For the common RGB refiner in Table 2, RIMAR scores 39.23 mIoU without refinement and 41.41 with it, versus 36.51 and 39.20 for PEARL. The paired RIMAR-PEARL differences are $+2.715$ points ($[2.549, 2.907]$) before refinement and $+2.203$ ($[1.960, 2.448]$) after it. The common interface does not assert identical native score calibration. For the native RIMAR stages, unrefined scores give $39.230/68.000$ mIoU/pAcc, Gaussian-only output gives $41.078/70.115$, and the complete blend, repooling, and robust updates give $41.394/70.643$. Gaussian-only and complete refinement take 4.35 and 20.11 ms/image on identical cached upstream fields: the latter adds 0.316 mIoU points and 15.76 ms/image, including the intervening blend and repooling.

Full RIMAR scores 41.394 mIoU versus 40.950 for the joint simplification. Its additional predictor latency is 7.53 ms/image, with paired 95% interval $[7.22, 7.83]$, under the measurement conditions stated in Sec. 4.3. Without RGB, feature distance improves mIoU by 2.667 points ($[2.552, 2.921]$); with RGB, its effect is -0.014 ($[-0.033, 0.004]$). The RGB-feature interaction is -2.681 ($[-2.937, -2.565]$).

Fixed aggregation counterpart. The fixed rule in Table 5 averages all stored patch states, retains the last class token, uses uniform log-probability pooling of native attention, balances that pool and BAFAT equally, averages each local/context branch over covering crops, and mixes the branches 1:1. Relative to fixing both rules, adaptive aggregation gains 4.377 mIoU points ($[4.130, 4.620]$) and 3.298 pAcc points ($[3.168, 3.427]$). With adaptive windows retained, the layer effect is $+4.199$ mIoU; with adaptive layers retained, the window effect is $+0.290$. Their factorial interaction is $+0.113$ ($[0.035, 0.166]$).

Table 10: Layer-policy controls with unchanged row-local attention consensus and downstream processing. The population policy is setting-specific and estimated from a disjoint unlabeled image bank; shuffled reports the mean of three fixed donor assignments. Values are equal-setting means (%).

Layer policy	Pre mIoU	Final mIoU	Final pAcc
Native per view	**39.230**	**41.394**	**70.643**
Fixed late layers	38.295	40.737	69.097
Population average	39.149	41.328	70.544
Mean shuffled	39.167	41.325	70.546

Native-minus-population and native-minus-mean-shuffled mIoU are $+0.066$ ($[0.026, 0.104]$) and $+0.069$ ($[0.034, 0.103]$), respectively. These paired intervals condition on the frozen calibration bank and donor maps. The population comparator uses extra unlabeled RGB images, whereas native RIMAR does not require that bank; none of these controls disables later adaptive semantic, window, or row-local attention operations.

The close substitutions in Table 11 are near-tied. Bidirectionality improves over the one-direction affine fit, but the aggregate data do not establish a material advantage of affine freedom over orthogonal or direct self-self alternatives. This qualification does not alter the complete-predictor comparison.

C.3 Sensitivity Across Input Conditions

Table 12: Sampling-density study: mean mIoU on the common reference support. Negative/positive amplitude concentrates samples near the boundary/center; ratio is the maximum-to-minimum density. Uniform BAFAT changes only its area weights.

a	Ratio	RIMAR	PEARL	NACLIP	Uniform BAFAT
-0.6	16.00	**30.16**	27.73	27.94	30.03
-0.3	3.45	**32.26**	30.53	30.33	32.16
0	1.00	33.82	32.24	32.04	**33.94**
0.3	3.45	**33.65**	31.91	31.90	33.63
0.6	16.00	31.44	29.77	30.06	**31.50**

The backbone comparison remains in main-text Table 6; the full resolution sweep appears in Tables 14-15. The paired-view studies vary perspective, equidistant-fisheye, and local ERP projections at three angular-cap diameters, or redistribute samples within a fixed square 90° perspective field. The circular-cap perspective reference in the projection study and the square reference in the sampling study are distinct. The "Label resampling" curve in Fig. 5 is an annotation-resampling control, not an upper bound on prediction agreement.

Table 12 complements the main-text view comparison in Fig. 5. RIMAR has the highest complete-predictor mIoU at every sampled amplitude, but the uniform-area BAFAT control is slightly higher at $a = 0$ and $a = 0.6$. Together with the feature-distance controls, this prevents attributing the complete predictor's view behavior to any single use of geometry.

C.4 Concise Evaluation Notes

The benchmark uses the frozen OpenAI CLIP ViT-B/16 checkpoint, the dataset class/alias mappings used by each reported setting, and an 80-template prompt ensemble. Perspective short sides are 336 except Cityscapes (560); panoramic and fisheye short sides are 640, with a 2048-pixel long-edge cap. VOC uses full-image inference; all other settings use 224×224 windows at stride 112. The score scale, posterior resolution, blend, heavy-tail scale, finite budgets, layer support, and semantic gate are inherited empirical choices; recovered records do not identify their original search procedure or development split. The headline predictor uses neither learned calibration nor an unlabeled calibration bank. Invalid or ignored pixels are excluded according to each benchmark's annotation convention. Intersections, unions, correct pixels, and valid pixels are summed within each physical split before

Table 11: Close alignment substitutions inside the otherwise unchanged predictor. Entries are mIoU (%); the final columns additionally give equal-setting mean mIoU and pAcc.

Panoramic and fisheye settings

Alignment	Wild	M3D	S2D	FDD	Wood
Self-self	52.66	**25.19**	38.57	28.73	**26.82**
Bidir. orthogonal	52.34	25.18	38.35	**28.80**	26.81
Forward affine	**53.06**	24.92	**38.69**	28.43	26.20
Bidir. affine	52.60	25.15	38.63	28.70	26.68

Perspective settings and overall means

Alignment	V21	PC60	Obj.	V20	PC59	Stuff	City	ADE	mIoU	pAcc
Self-self	68.38	37.67	38.89	88.58	41.43	26.90	**43.08**	20.38	41.328	**70.666**
Bidir. orthogonal	**68.73**	**37.81**	39.09	88.69	**41.62**	**27.11**	43.03	**20.45**	41.386	70.644
Forward affine	68.07	37.76	39.18	**88.86**	41.59	27.05	42.42	20.25	41.268	70.435
Bidir. affine	68.62	37.75	**39.20**	88.75	41.59	27.08	42.94	20.44	**41.394**	70.643

metrics are computed. Stanford2D3D's three physical area groups are averaged equally into one logical score; the 13 logical settings are then averaged equally.

Tables 1 and 8 report the broad benchmark at display precision. The fusion, common-refiner, guidance, and paired-view studies are separate controlled evaluations. Their intervals use paired resampling across the same images or source panoramas and preserve shared multiplicities across compared variants. Mechanism contrasts use 2,000 paired image-bootstrap replicates with shared multiplicities; the view studies use 1,000 paired source-panorama replicates. The intervals are pointwise and conditional on the evaluated populations, not a measure of generalization across independent buildings or cameras. Random components used fixed seeds where recorded.

Full-image VOC bypasses overlapping-window integration; positive-extent inference uses 224×224 crops at stride 112 and executes the declared integration even if only one crop covers an image. RIMAR's refiner operates on a 112×112 posterior grid, restores bilinearly, and decodes at original resolution. Ground-truth labels are used for evaluation-metric calculation and the disclosed qualitative-example selection, but not for adaptive inference weights or graph construction.

C.5 Full Resolution Sweep and Component Ablation

Tables 14-15 report the full 13-setting resolution sweep. It is separate from the headline protocol, which uses short sides of 640 for panoramic/fisheye images and 560 for Cityscapes. RIMAR leads the mean at 224, 336, and 448 pixels: mIoU is $37.9/40.5/40.5$ versus the strongest baseline's $34.8/37.0/37.6$, and pAcc is $66.9/70.1/70.3$ versus $62.6/65.4/66.3$. Improvement largely saturates beyond 336.

Broad component ablation. Table 13 reports the broad component ablation separately from the narrower controls in Secs. 4.3-4.4. BAFAT removal gives the largest loss; the flat-metric and no-MAP variants also reduce both metrics, while removing either MAP stage gives smaller losses.

Table 13: Ablation study averaged by image domain. Values are mIoU and pAcc (%), with domain averages over the panoramic, fisheye, and perspective datasets and Avg. over all 13 datasets.

Variant	mIoU				pAcc			
	Pano.	Fish.	Persp.	Avg.	Pano.	Fish.	Persp.	Avg.
RIMAR	38.8	27.7	45.8	41.4	68.4	71.9	71.2	70.6
w/o BAFAT	10.4	5.5	14.2	12.0	26.9	26.9	39.7	34.8
w/ flat metric	27.9	19.9	40.3	34.3	58.5	64.3	68.0	65.2
w/o MAP regularization	37.5	25.7	43.3	39.2	65.4	68.3	68.9	68.0
w/o Gaussian model	38.5	27.3	44.9	40.7	67.2	70.6	70.3	69.7
w/o Heavy-tail model	38.4	26.9	44.8	40.6	67.0	70.1	70.2	69.4

Table 14: Resolution sensitivity in mIoU (%) across 13 datasets. Avg. is the arithmetic mean over datasets.

Res.	Method	Panoramic			Fisheye		Perspective								Avg.
		Wild	M3D	S2D	FDD	Wood	V21	PC60	Obj.	V20	PC59	Stuff	City	ADE	
224	NACLIP	45.8	20.7	**29.7**	18.0	16.7	62.4	34.1	35.7	83.6	37.4	24.8	25.8	17.4	34.8
	PEARL	42.4	18.9	26.0	15.7	14.5	59.1	33.1	36.1	86.1	36.4	24.9	24.5	17.7	33.5
	RIMAR	**50.0**	**21.5**	29.6	**23.9**	**18.3**	**65.0**	**37.3**	**39.2**	**87.8**	**40.9**	**26.7**	**33.7**	**18.9**	**37.9**
336	NACLIP	47.2	23.1	35.1	22.1	20.2	64.1	35.0	36.2	83.0	38.4	25.7	32.4	19.1	37.0
	PEARL	43.7	21.9	34.3	20.5	17.3	64.1	35.1	37.3	86.9	38.6	26.3	31.5	19.4	36.7
	RIMAR	**52.4**	**24.4**	**38.1**	**26.7**	**23.1**	**68.6**	**37.8**	**39.2**	**88.8**	**41.6**	**27.1**	**38.9**	**20.4**	**40.5**
448	NACLIP	46.8	24.3	36.9	24.5	22.8	60.8	34.4	34.8	77.6	37.6	25.2	36.1	19.2	37.0
	PEARL	43.6	23.4	36.7	23.6	20.2	64.1	35.2	37.0	84.5	38.7	26.2	35.4	19.6	37.6
	RIMAR	**52.4**	**25.1**	**39.1**	**27.8**	**25.3**	**68.0**	**36.9**	**37.2**	**86.7**	**40.3**	**26.3**	**41.2**	**20.1**	**40.5**

Table 15: Resolution sensitivity in pixel accuracy (pAcc, %) across the same 13 datasets. Avg. is the arithmetic mean over datasets.

Res.	Method	Panoramic			Fisheye		Perspective								Avg.
		Wild	M3D	S2D	FDD	Wood	V21	PC60	Obj.	V20	PC59	Stuff	City	ADE	
224	NACLIP	73.2	48.0	**57.3**	69.2	51.2	88.6	51.6	75.7	91.9	61.3	40.4	61.7	44.0	62.6
	PEARL	72.0	43.8	51.3	65.5	48.9	86.0	50.8	73.9	92.9	60.1	40.9	60.7	42.6	60.7
	RIMAR	**77.5**	**50.2**	54.4	**77.8**	**53.8**	**89.7**	**56.4**	**79.4**	**94.0**	**65.9**	**44.8**	**76.7**	**49.5**	**66.9**
336	NACLIP	72.9	52.3	65.4	74.5	53.6	89.3	53.5	77.1	90.5	62.8	41.7	68.8	48.1	65.4
	PEARL	71.8	49.0	66.1	72.8	50.4	88.5	53.5	75.8	93.3	62.9	42.9	68.4	46.9	64.8
	RIMAR	**78.0**	**55.2**	**71.6**	**80.2**	**59.0**	**91.2**	**57.7**	**80.3**	**94.4**	**66.9**	**45.6**	**79.0**	**52.3**	**70.1**
448	NACLIP	70.6	54.8	66.8	76.9	57.2	89.2	53.7	77.6	89.0	62.4	41.4	72.2	49.5	66.3
	PEARL	70.0	51.8	68.9	76.0	55.3	88.9	54.3	76.7	92.0	63.2	43.0	72.3	48.5	66.2
	RIMAR	**76.9**	**57.0**	**71.6**	**80.9**	**61.2**	**91.2**	**57.2**	**80.0**	**93.2**	**66.1**	**44.7**	**80.4**	**53.0**	**70.3**