RETAX-SEG: VISUALLY GROUNDED REVERSIBLE TAXONOMY TRANSPORT FOR IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing taxonomy-aware segmenters reconcile incompatible label names but leave a fundamental ambiguity: a dataset label can be a policy-dependent union of visual concepts, yet its mapping to a unified vocabulary remains global and one-way. A fixed mapping cannot identify the constituent present in a region. A language-derived mapping imports name semantics without observing that region. A learned latent vocabulary can also collapse despite correct final-label predictions. ReTax-Seg addresses these failures jointly. A visual decoder predicts 39 fine-grained semantic atoms, and an end-to-end transport field compiles them into native labels with a different atom–class distribution at every spatial location. Class names supply only a learned-strength CLIP prior, while supervised visual features and reliability-gated MobileSAM geometry determine the local correction. Reverse transport requires predicted native labels to reconstruct the visual atom distribution, making semantic preservation an explicit training objective. To our knowledge, ReTax-Seg first combines pixel-conditioned class–atom transport, visual-over-language arbitration, and bidirectional taxonomy consistency in one differentiable compiler. Six independent checkpoints average 73.63% test mIoU. One jointly trained six-dataset checkpoint, ReTax-Seg-C, shares the atom representation and reaches 74.05%, outperforming the compared methods on every dataset.

Keywords: Semantic segmentation, taxonomy alignment, visual grounding, dynamic transport, cycle consistency.

1 INTRODUCTION

Semantic segmentation labels are visual categories shaped by annotation policy: datasets merge, split, omit, or rename the same visible concepts. Existing work reconciles taxonomies through composite vocabularies (Lambert et al., 2020, Zhang et al., 2025c, Yang et al., 2025, Gao et al., 2025, Fu et al., 2026) and unions of universal classes (Bevandić et al., 2022, Bevandić & Šegvić, 2022). Other approaches use language embeddings (Zhou et al., 2023, Zhang et al., 2025b,a) or learned dynamic alignment (Ma et al., 2024, Sun et al., 2025). These methods leave one local ambiguity unresolved: *which constituent of a coarse or policy-dependent class is supported by the pixels at this location?*

A coarse *vehicle* mask may denote a car, bus, motorcycle, or bicycle, while *construction* may absorb buildings, walls, fences, or poles. A static mapping assigns every occurrence the same mixture. It records what a label may contain, not what the current region visibly contains. Fixed taxonomies therefore cannot disambiguate locally. Language alignment cannot see the deciding pixels. Freely learned latent mappings can collapse under forward label supervision. A useful solution must be *local*, *visually grounded*, and *semantically recoverable* (Figure 1).

We meet these requirements with one factorization:

$$\text{image} \longrightarrow \text{semantic atoms} \longrightarrow \text{region-dependent dataset labels.}$$

The visual network predicts 39 fine-grained atoms, and a region-conditioned transport field compiles them into native labels. A human-auditable base relation and a globally scaled frozen CLIP prior (Radford et al., 2021, Li et al., 2026, Kang et al., 2025) are corrected by local visual and geometric evidence. Text therefore cannot produce a mask independently. Because final-label supervision alone leaves the atom decomposition underdetermined, the same transport maps predicted labels back to

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.

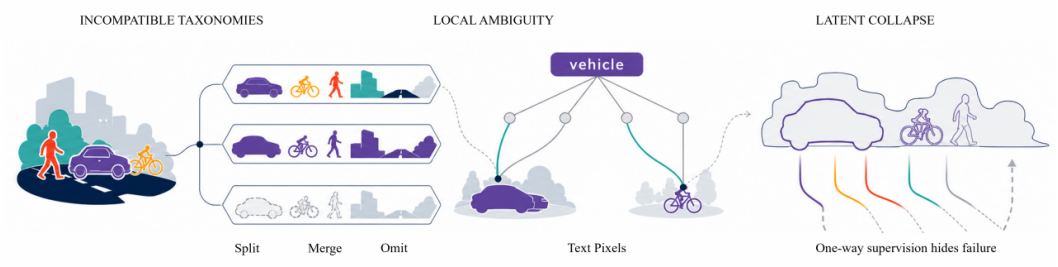

Figure 1: **Three unresolved challenges in taxonomy-aware segmentation.** Left: identical visual concepts can be split, merged, or omitted by different annotation policies. Center: a coarse label such as *vehicle* is compatible with several semantic atoms, but the correct relation changes with local visual evidence, which a class name alone cannot observe. Right: forward-only supervision can preserve an apparently correct coarse mask even when distinct atom channels become indistinguishable. A useful compiler must therefore be local, visually grounded, and semantically recoverable.

atoms (Yang et al., 2025, Chen et al., 2025). Cycle consistency requires the forward atom distribution to remain recoverable.

We evaluate six independent checkpoints and one cooperative checkpoint, separating per-taxonomy trainability from cross-dataset atom sharing.

Our contributions are:

- We introduce a visually grounded reversible taxonomy compiler whose atom–class transport varies spatially and must recover its visual atoms.
- We separate authority between priors and observations. Class names provide a soft global prior. Supervised pixels and reliability-gated geometry determine local transport. Bidirectional consistency discourages atom collapse.
- Across six datasets and six metrics, independent training averages 73.63% test mIoU, while cooperative training raises the mean to 74.05% using one shared checkpoint.

2 Related Work

Semantic segmentation. Dense predictors evolved from fully convolutional networks (Long et al., 2015) through atrous models (Chen et al., 2017) to transformer encoders such as SegFormer (Xie et al., 2021). They normally end in a fixed classifier. ReTax-Seg instead places an atom-to-label compiler above a MiT-B5 backbone.

Taxonomy reconciliation. MSeg relabels ambiguous examples into a composite taxonomy (Lambert et al., 2020). Universal-taxonomy methods retain native annotations as unions of finer classes (Bevandić et al., 2022, Bevandić & Šegvić, 2022). AutoUniSeg learns a language-initialized unified space (Ma et al., 2024). These methods discover a largely global relation between label systems. We instead ask which constituent a particular region supports and whether that local decision remains recoverable.

Language-guided and dynamic taxonomy alignment. LMSeg redirects text-aligned segment queries into a requested taxonomy (Zhou et al., 2023), while DynAlign combines semantic knowledge, foundation-model masks, and scene adaptation for annotation-free target transfer (Sun et al., 2025). Their alignment remains forward-only. We learn dense local transport, restrict text to a soft prior, and reconstruct the visual atoms through the same transport. The six independent models share the atom vocabulary and compiler design.

Language and geometric priors. CLIP embeds visual and textual concepts (Radford et al., 2021). SAM and MobileSAM provide class-agnostic region cues (Kirillov et al., 2023, Zhang et al., 2023a, An et al., 2025, Chen et al., 2026). We use neither as an oracle: CLIP adds only an offline scaled

bias, MobileSAM is frozen behind a learned reliability gate, and native masks provide all pixel-level semantics.

3 METHOD

3.1 FROM STATIC CLASSIFICATION TO SEMANTIC COMPILATION

Throughout, d, x, r, c, and z index a dataset, image, decoder location, native class, and atom. The corresponding sets are $\mathcal{C}_d$ and $\mathcal{Z}$, with $K = |\mathcal{Z}| = 39$. Here $|\cdot|$ is cardinality, $\mathbb{R}^n$ is real n-space, and Δ^{n-1} is a probability simplex.

ReTax-Seg trains one checkpoint per d. ReTax-Seg-C uses C for *cooperation* and jointly trains all six datasets. It shares the backbone, atom decoder, region module, and atom factors. Base transport, class factors, and the gated 1×1 adapter remain dataset-specific.

The network f_θ has learnable parameters θ. At r, it outputs the atom-logit vector $a(x, r) \in \mathbb{R}^K$, whose zth component is a_z:

$$a(x, r) = f_\theta(x)_r, \qquad p_\theta(z \mid x, r) = \mathrm{softmax}_z\, a_z(x, r). \tag{1}$$

A compiler $T_d(c \mid z, x, r)$ assigns atom z to native class c at each location. The resulting native-class posterior is

$$p(y_d = c \mid x, r) = \sum_{z \in \mathcal{Z}} T_d(c \mid z, x, r) p_\theta(z \mid x, r). \tag{2}$$

Here $\mathrm{softmax}_z$ normalizes over atoms. Each $T_d(\cdot \mid z, x, r)$ is a distribution over $\mathcal{C}_d$. It depends on the local region. The native mask directly supervises Equation 2. No atom-level ground truth or altered annotation is used.

3.2 ATOMIC VISUAL DECODER

MiT-B5 extracts four feature scales with 64, 128, 320, and 512 channels. Each feature is projected to 128 channels, resized to one-eighth input resolution, and fused by a 3×3 convolution, group normalization, GELU, and spatial dropout. A 1×1 layer produces the 39 atom logits. The atom vocabulary is fixed across all independent experiments, while the active compiler targets the native classes of the current dataset.

3.3 RELIABILITY-GATED REGION REPRESENTATION

Dynamic compilation needs a compact description of local content and boundaries. A learned visual branch extracts $S = 8$ region factors $R_v(h)$ from the fused decoder feature h. In parallel, a frozen MobileSAM image encoder processes an internal 1024×1024 view. It produces a class-agnostic feature tensor s. We combine both branches as

$$\rho(x, r) = \sigma(g_\rho(h_r \parallel \phi_s(s)_r)), \tag{3}$$

$$q(x, r) = \tanh(R_v(h)_r + \rho(x, r) R_s(\phi_s(s))_r), \tag{4}$$

Here $h_r \in \mathbb{R}^{128}$ is the fused visual vector. The learned map ϕ_s projects the SAM tensor. The maps R_v and R_s project to S channels. The symbol $\parallel$ is concatenation, and g_ρ is the gate network. The elementwise functions σ and $\tanh$ are sigmoid and hyperbolic tangent. The scalar $\rho \in (0, 1)$ is reliability, and $q \in \mathbb{R}^S$ is the region state. The visual branch is always active.

3.4 TEXT-INFORMED REGION-CONDITIONED TRANSPORT

We initialize a base matrix $A_d \in \mathbb{R}^{|\mathcal{C}_d| \times K}$ from a sparse, human-inspectable compatibility specification. Compatible entries begin at $+2.5$, and incompatible entries begin at -2.5. All entries remain trainable.

To use label names without allowing language to decide the segmentation, we precompute frozen CLIP embeddings for short class descriptions and atom descriptions. Centered, standardized cosine

affinity gives $S^{\text{text}}_{d,c,z}$. This matrix is spatially constant and has no access to the image. Its learned scale is $\alpha = \text{softplus}(\tilde{\alpha}) \geq 0$, where $\tilde{\alpha}$ is an unconstrained scalar.

Image evidence supplies the spatial correction. Dataset-class factors $U_d \in \mathbb{R}^{|\mathcal{C}_d| \times S}$ and atom factors $V \in \mathbb{R}^{K \times S}$ interact with $q(x, r)$. Here u indexes one of the S low-rank factors. The unnormalized transport logit ℓ and its normalized probability T_d are

$$\ell_{d,c,z}(x, r) = A_{d,c,z} + \alpha S^{\text{text}}_{d,c,z} + \sum_{u=1}^{S} q_u(x, r) U_{d,c,u} V_{z,u}, \tag{5}$$

$$T_d(c \mid z, x, r) = \text{softmax}_c \, \ell_{d,c,z}(x, r). \tag{6}$$

Text proposes a global preference. The trainable base can correct naming or policy mismatch. The region term changes the relation locally. The operator softmax_c makes each transport column sum to one. Outputs still depend on the visually predicted atoms in Equation 2. No language model runs during training or inference.

The same construction has a compact matrix form. Let $\pi_r = [p_\theta(z \mid x, r)]_{z \in \mathcal{Z}} \in \Delta^{K-1}$ be the atom posterior vector. Let $p_{d,r} = [p(y_d = c \mid x, r)]_{c \in \mathcal{C}_d}$ be the native posterior vector. We use $q_r = q(x, r)$ and $T_{d,r}(c, z) = T_d(c \mid z, x, r)$ as compact notation. Then

$$L_{d,r} = A_d + \alpha S^{\text{text}}_d + U_d \, \text{Diag}(q_r) V^\top, \qquad p_{d,r} = T_{d,r} \pi_r, \quad T_{d,r} = \text{softmax}_{\text{col}}(L_{d,r}). \tag{7}$$

Here $L_{d,r}$ is the logit matrix. The operator Diag forms a diagonal matrix, and $\top$ denotes transpose. Column softmax normalizes each atom over native classes. The conditional correction has rank at most $S = 8$, although A_d remains a full trainable matrix. Thus, the model can change a complete class–atom table at every location while forcing those changes to lie in a shared, compact set of semantic directions. This separates global taxonomy knowledge from the spatial degrees of freedom needed to resolve local ambiguity.

Native supervision reaches both sides of the compiler. For a valid pixel, let $y_r \in \mathcal{C}_d$ be its target class. The posterior responsibility of atom z is

$$\gamma_{r,z} = \frac{T_{d,r}(y_r \mid z) \pi_{r,z}}{\sum_j T_{d,r}(y_r \mid j) \pi_{r,j}}. \tag{8}$$

Here j is a dummy atom index. Let $\mathcal{L}_r = -\log p(y_d = y_r \mid x, r)$ be the pixel loss. We abbreviate $a_{r,z} = a_z(x, r)$ and $\ell_{r,c,z} = \ell_{d,c,z}(x, r)$. Then

$$\frac{\partial \mathcal{L}_r}{\partial a_{r,z}} = \pi_{r,z} - \gamma_{r,z}, \qquad \frac{\partial \mathcal{L}_r}{\partial \ell_{r,c,z}} = \gamma_{r,z} \big(T_{d,r}(c \mid z) - \mathbb{1}[c = y_r] \big). \tag{9}$$

The indicator $\mathbb{1}[c = y_r]$ equals one when $c = y_r$ and zero otherwise. Hence, the native mask does not assign an atom by fiat. It induces a soft responsibility through agreement between visual evidence and the current transport. High-responsibility atoms receive the strongest transport update. The atom head is pulled toward the same posterior explanation.

3.5 Why the mapping must vary spatially

A fixed compiler would remove the final term in Equation 6. It must then encode a broad class, for example *construction*, as the same global distribution over building, wall, fence, pole, and temporary-structure atoms at every pixel. The dynamic compiler instead uses $q(x, r)$ to emphasize the locally supported relation: building-like atoms in a facade region, pole-like atoms on a thin vertical structure, and other construction atoms near temporary works. This is not a change to the dataset's class definition. It is a change to the fine-grained explanation used to realize that definition from the current pixels.

3.6 Bidirectional semantic constraint

Coarse masks do not uniquely determine an atom decomposition. Without an additional constraint, several atom channels can become interchangeable or collapse into arbitrary latent coordinates. We transport the predicted label distribution back through the same spatial compiler:

$$\hat{p}(z \mid x, r) = \frac{\sum_c p(y_d = c \mid x, r) T_d(c \mid z, x, r)}{\sum_{z'} \sum_c p(y_d = c \mid x, r) T_d(c \mid z', x, r)}. \tag{10}$$

The cycle loss is

$$\mathcal{L}_{\text{cycle}} = D_{\text{KL}}(\text{stopgrad}[p_\theta(z \mid x, r)] \,\|\, \hat{p}(z \mid x, r)) \,. \tag{11}$$

Here $\hat{p}$ is the reconstructed atom posterior, and z' is a dummy atom index used for normalization. For atom distributions P and Q, the divergence is $D_{\text{KL}}(P\|Q) = \sum_z P_z \log(P_z/Q_z)$. The stopgrad operator returns its argument but blocks its gradient. Detaching the forward target requires the reverse path to explain the atom distribution rather than allowing both distributions to drift toward a trivial agreement. The compiler is therefore trained in both directions: atoms must predict labels, and the resulting labels must preserve enough information to reconstruct the atoms.

More explicitly, write $\mathcal{N}(v) = v/(\mathbf{1}^\top v)$. The vector $\mathbf{1}$ contains ones and has the same length as v. Suppressing the dataset and image arguments gives

$$\hat{\pi}_r = \mathcal{N}\big(T_r^\top T_r \pi_r\big) \,. \tag{12}$$

The vector $\hat{\pi}_r$ collects the reconstructed probabilities $\hat{p}(z \mid x, r)$ over all atoms. The Gram entry $[T_r^\top T_r]_{z,z'}$ is the overlap between the native-class distributions assigned to atoms z and z'. Identical transport columns therefore become indistinguishable under the cycle. Since many datasets have fewer than 39 classes, global inversion of the whole atom simplex is impossible. Equation 11 instead enforces recoverability on the atom distributions actually occupied by the images. This is a local identifiability constraint, not a claim of a bijection between taxonomies.

3.7 Training objective

The complete loss is

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{seg}} &+ 0.05\mathcal{L}_{\text{cycle}} + 0.01\mathcal{L}_{\text{transport-ent}} + 0.02\mathcal{L}_{\text{region-TV}} \\ &+ 0.05\mathcal{L}_{\text{boundary}} + 0.02\mathcal{L}_{\text{region-boundary}}.\end{aligned} \tag{13}$$

$\mathcal{L}_{\text{seg}}$ is pixel-wise negative log likelihood with ignore index 255. Transport entropy discourages diffuse atom-to-class assignments. Total variation stabilizes the region field. Output-boundary and region-boundary losses align probability transitions and soft region transitions with native label boundaries. The symbol $\mathcal{L}$ denotes the scalar training objective. Each numeric coefficient in Equation 13 is a fixed loss weight. The matrix derivation, exact gradient flow, normalized cycle operator, and finite-difference definitions of all regularizers are given in Appendix C.

4 Experimental Protocol

4.1 Datasets and actual input modality

We evaluate six independent models and one cooperatively trained model. AeroScapes contains aerial imagery (Nigam et al., 2018). VDD contains varied drone scenes (Cai et al., 2025). MFNet contributes its processed three-channel visible images (Ha et al., 2017). FMB contributes its `Visible` branch (Liu et al., 2022). Cityscapes provides fine urban annotations (Cordts et al., 2016). SUIM contains underwater RGB imagery (Islam et al., 2020). The present pipeline reads three-channel RGB/visible inputs for every experiment.

Evaluation splits. All the datasets were divided into training set, validation set and test set in a ratio of 8:1:1. We accumulate one confusion matrix over every valid pixel and report overall accuracy (aAcc), macro recall (mAcc), macro precision (mPrec), macro F1/Dice, mIoU, and frequency-weighted IoU (fwIoU).

4.2 Implementation details

Independent models use 40k iterations and select the highest-validation- mIoU checkpoint. ReTax-Seg-C uses one 40k-iteration run, round-robin mini-batches, and mean validation mIoU for selection. All runs use MiT-B5, 39 semantic atoms, a 128-channel decoder, rank-8 transport, 512×512 crops, and a frozen MobileSAM encoder operating internally at 1024×1024. We optimize for 40k iterations with validation every 4k, batch size 2, mixed precision, and AdamW (base learning rate 6×10^{-5}, weight decay 0.01, backbone multiplier 0.2). A 1.5k-iteration warmup precedes polynomial decay

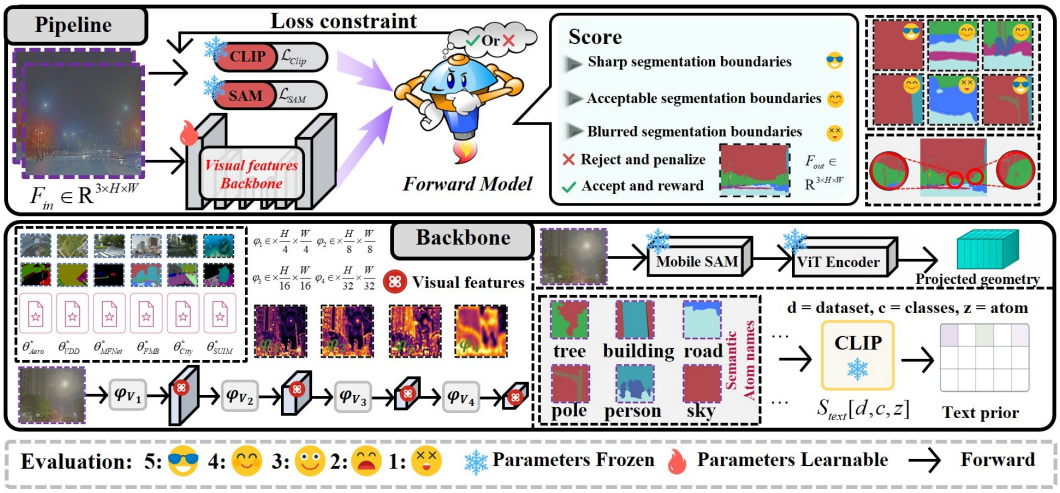

Figure 2: **Overview of ReTax-Seg.** The upper panel summarizes the training pipeline. The visual backbone and forward model are learnable. CLIP and SAM remain frozen. Boundary quality scores the segmentation outputs. The lower panel details the multiscale visual backbone, the frozen MobileSAM geometry branch, representative semantic atoms, and the frozen CLIP text-prior construction. The symbol θ_d^* denotes the weights associated with dataset d. The star marks the selected trained checkpoint. For example, θ_{City}^* represents the Cityscapes-specific weights. Snowflakes and flames indicate frozen and learnable components, respectively.

with power 0.9. Gradients are clipped at 5.0 and the random seed is 0. MiT-B5 is initialized from ImageNet weights, whereas MobileSAM and the offline CLIP text bank remain frozen. Training uses only native masks. No online language model is used in training or inference.

5 Results

5.1 Validation performance

The best validation checkpoints occur at 40k iterations for AeroScapes (83.61% mIoU), VDD (81.84%), MFNet (71.91%), and Cityscapes (72.67%). FMB peaks at 32k iterations (71.95%), and SUIM peaks at 36k (68.91%). Their unweighted mean is 75.15% mIoU. The earlier FMB and SUIM optima support validation-based rather than final-iteration checkpoint selection.

5.2 Six-dataset comparison

Table 1 compares 20 archived alternatives. They include efficient networks (Wu et al., 2020, Zhao et al., 2018, Wan et al., 2023, Zhang et al., 2022b), HRNet, and transformer models (Wang et al., 2020, Xie et al., 2021, 2025). The comparison also covers DeepLabV2 variants (Chen et al., 2017, He et al., 2016, Zhang et al., 2022a, Xie et al., 2017) and two SAM-based models (Xiong et al., 2026, Ke et al., 2023). The text-guided group contains LSeg, SemiVL, and PartCATSeg (Li et al., 2022, Hoyer et al., 2024, Choi et al., 2025). The baselines and ReTax-Seg use one checkpoint per native taxonomy. ReTax-Seg-C instead denotes one checkpoint jointly optimized on all six training sets: it shares the visual and atomic representation, while dataset identity selects the native transport and a gated adapter. The row therefore measures cross-domain sharing without merging incompatible output spaces. All methods use 512×512 inputs and validation-selected checkpoints. Metrics are recomputed from complete predictions under the common protocol. The main table reports mF1 and mIoU. The metrics aAcc, mAcc, mPrec, and fwIoU are retained in Appendix Tables 5 and 6 to keep the comparison legible at the document's normal table font size.

Independent ReTax-Seg averages 73.63% mIoU. ReTax-Seg-C improves all six dataset values and raises the mean to 74.05% (+0.42 points). Its mean mF1 also rises from 83.46% to 84.11%. This

Table 1: mF1 and mIoU comparison on all datasets (%). Both metrics are computed from complete confusion matrices. Test is used when labeled test masks exist, otherwise validation. Column bests are bold. The metrics aAcc, mAcc, mPrec, and fwIoU are reported in Appendix Tables 5 and 6. The suffix C in ReTax-Seg-C denotes cooperation, namely joint training on all six datasets.

Method	**AeroScapes**		**VDD**		**MFNet**		**FMB**		**Cityscapes**		**SUIM**	
	mF1	mIoU	mF1	mIoU	mF1	mIoU	mF1	mIoU	mF1	mIoU	mF1	mIoU
CGNet	49.07	38.78	70.08	56.55	54.60	41.00	47.61	37.76	42.82	32.68	43.90	32.02
HRNet	80.49	68.88	89.70	81.86	82.01	70.42	61.23	49.81	68.24	55.12	58.38	46.88
ICNet	76.20	65.46	82.44	72.00	64.31	50.29	48.57	38.57	46.15	35.66	41.06	29.90
SegFormer-B3	75.21	64.76	81.21	70.48	73.53	59.59	49.10	39.50	59.87	46.98	56.54	43.41
SegFormer-B4	73.57	63.07	81.81	71.31	73.55	59.60	49.86	40.12	58.91	46.18	57.18	44.41
R101-SegFormer	59.63	46.99	81.81	70.63	66.45	52.43	51.10	40.59	57.92	45.64	40.83	30.00
R50-DLV2	63.29	51.00	79.60	68.26	65.58	51.68	51.30	40.81	60.87	47.73	60.59	47.04
ResNeSt-50-DLV2	82.74	72.43	88.67	80.28	80.60	68.68	57.20	46.55	68.96	55.46	61.64	48.10
ResNeSt-101-DLV2	84.31	74.97	89.59	81.72	81.54	69.88	63.35	51.55	72.55	59.47	66.58	52.43
ResNeSt-200-DLV2	87.59	79.06	90.07	82.43	82.16	70.69	63.62	52.14	71.96	58.70	68.47	55.00
SeaFormer	41.31	32.88	59.06	47.67	40.14	30.27	23.13	17.32	38.58	30.04	44.01	32.56
TopFormer	34.36	27.59	69.15	57.35	47.86	35.82	41.54	32.47	36.84	28.38	42.60	31.48
ResNeXt-50-DLV2	84.91	74.89	88.88	80.65	80.81	68.84	64.51	52.39	74.12	60.99	63.26	49.93
ResNeXt-101-DLV2	82.45	72.01	88.63	80.22	80.90	69.08	65.39	53.48	74.47	61.76	63.54	50.15
MaSSFormer	80.54	70.53	84.86	74.48	71.58	59.26	65.58	53.32	23.04	18.42	59.26	46.80
SAM2-UNet	83.94	74.08	87.83	79.24	81.45	69.95	72.51	60.05	77.49	65.10	79.30	68.99
HQ-SAM	73.39	61.24	81.58	70.85	66.92	53.73	54.67	44.87	62.33	49.88	63.45	50.39
LSeg	77.74	67.81	81.80	72.32	73.77	62.13	65.34	52.93	68.72	55.87	32.23	27.41
SemiVL	79.69	69.13	90.18	82.67	80.72	68.88	63.88	51.83	83.07	72.50	80.07	68.50
PartCATSeg	78.05	66.95	75.74	65.03	77.94	65.79	61.81	49.59	76.95	65.08	80.41	69.42
ReTax-Seg	**88.28**	**80.34**	**91.35**	**84.50**	**82.97**	**71.91**	**72.69**	**60.64**	**83.18**	**72.67**	**82.30**	**71.74**
ReTax-Seg-C	**88.76**	**80.92**	**92.03**	**84.97**	**83.55**	**72.38**	**73.48**	**61.27**	**83.98**	**72.75**	**82.86**	**72.03**

consistent change indicates positive cross-domain transfer under cooperative training. SAM2-UNet and HQ-SAM average 69.57% and 55.16% mIoU, respectively.

Table 2 adds inference cost. ReTax-Seg retains the highest mean mIoU at 50.34 ms (19.87 FPS) on a V100, slightly faster than SAM2-UNet while improving mean mIoU by 4.06 points.

Figure 3 complements the aggregate metrics with boundary- and object-level examples from five conventional and two text-guided comparators.

The complete-split table supports system-level comparison, while the selected crops expose local failure modes. Component effects are isolated below.

5.3 Cross-dataset generalization

We evaluate nine representative methods under strict zero-shot transfer. For each source dataset, its best checkpoint is tested on the other five datasets without target updates. Conventional native heads are converted once by a fixed class map maximizing definition overlap in the shared 39-atom vocabulary. ReTax-Seg instead compiles its visual atoms directly into the requested target taxonomy. Table 7 averages the five foreign-source results for each target.

Across all 30 directed transfers, ReTax-Seg obtains 36.13% aAcc, 17.79% mAcc, and 23.82% mPrec. It also obtains 12.42% mF1, 9.45% mIoU, and 27.93% fwIoU. The strongest comparator is HQ-SAM at 9.07% mean mIoU. The corresponding value for ReTax-Seg is 9.45%.

Table 2: Accuracy and inference cost at 512×512, batch size 1. Mean mIoU averages the six datasets. GPU results use one V100. CPU results use eight Intel Core Ultra 7 270K Plus threads.

Method	mIoU	Params.		FLOPs	State	Peak	Time		FPS
	(%)	Total	Train.	(G)	(MB)	(MB)	GPU ms	CPU ms	GPU
CGNet	39.80	**0.50**	**0.50**	**3.43**	**1.93**	**55.96**	**3.83**	**22.77**	**261.29**
HRNet	62.16	9.64	9.64	18.41	36.91	93.54	11.19	89.69	89.36
ICNet	48.65	47.83	47.83	15.38	182.68	395.30	6.03	57.58	165.97
SegFormer-B3	54.12	47.24	47.24	71.52	180.20	680.47	26.99	278.08	37.05
SegFormer-B4	54.11	64.01	64.01	85.59	244.18	744.03	34.61	331.07	28.90
ResNeSt-200-DLV2	66.34	68.71	68.71	272.40	262.93	403.99	70.67	578.25	14.15
SAM2-UNet	69.57	216.42	4.27	215.95	825.62	967.65	53.25	527.84	18.78
HQ-SAM	55.16	88.21	0.93	104.08	336.50	521.67	26.13	430.52	38.27
LSeg	56.41	260.33	260.33	125.68	753.11	1139.98	21.79	478.16	45.89
SemiVL	68.92	89.05	31.35	25.03	339.69	499.93	40.38	413.95	24.76
PartCATSeg	63.64	177.75	26.77	105.29	679.96	883.33	145.77	564.78	6.86
ReTax-Seg	**73.63**	88.41	82.34	99.25	343.78	592.36	50.34	496.63	19.87

GPU and CPU runs use FP32, except LSeg's checkpoint-native mixed-precision CLIP branch. Times exclude cached text setup, preprocessing, and I/O. One multiply–accumulate counts as one FLOP. State includes parameters and buffers at evaluation precision. Peak is allocated GPU memory.



Figure 3: **Qualitative comparison with two error-focused zooms per dataset.** Columns show the input, ground truth, five conventional models, SemiVL, PartCATSeg, and ReTax-Seg. The zooms are diagnostic regions where ReTax-Seg has higher local pixel accuracy than every displayed comparator. Table 1 remains the unbiased complete-split comparison.

5.4 CONTROLLED COMPONENT ABLATION

Table 3: Controlled ablation on the FMB and Aeroscapes. All variants use the matched backbone, data, crop, schedule, and seed.

	FMB			Aeroscapes		
Variant	mAcc	mF1/Dice	mIoU	mAcc	mF1/Dice	mIoU
Full dynamic ReTax-Seg	**68.74**	**72.69**	**60.64**	**87.51**	**88.28**	**80.34**
Static class–atom mapping	65.77	70.95	58.27	85.75	85.67	79.55
w/o CLIP affinity	66.75	71.72	59.27	83.71	84.60	78.45
w/o cycle consistency	65.08	70.11	59.51	86.99	86.62	79.49
w/o MobileSAM	66.25	69.36	57.98	85.81	86.63	79.01

Under matched settings, the full model is best on both datasets. Removing MobileSAM causes the largest FMB mIoU drop (2.66 points), while removing CLIP causes the largest AeroScapes drop (1.89 points). The cycle ablation reduces FMB mAcc by 3.66 points, confirming the complementary roles of local geometry, semantic initialization, and latent recovery.

Figure 4 compares boundaries and small objects across all six evaluation domains.



Figure 4: **Six focused boundary and small-object comparisons across the evaluation domains.** Rows follow AeroScapes, VDD, MFNet, FMB, Cityscapes, and SUIM. The leftmost panel provides the full-image context and marks the selected region. Subsequent columns show the RGB crop, ground truth (GT), CGNet, HRNet, ICNet, SegFormer-B3, SegFormer-B4, R101-SegFormer, R50-DLV2, ResNeSt-101, SeaFormer, and ReTax-Seg. Insets enlarge the target structure, and yellow borders identify the ReTax-Seg predictions.

6 CONCLUSION

We introduced ReTax-Seg, combining pixel-conditioned transport with reverse semantic preservation. Six independent models average 73.63% test mIoU, and the jointly trained ReTax-Seg-C reaches 74.05%, showing that a shared atom space can support several native taxonomies. Ablations confirm complementary benefits from dynamic transport, CLIP initialization, gated MobileSAM geometry, and cycle consistency. The current model still requires dataset identity and a predefined atom vocabulary, while bidirectional transport adds training cost. Future work will automate atom discovery and remove dataset routing.

AI use statement. This study did not utilize generative artificial intelligence tools for assistance. The authors reviewed and verified all evidence. They are fully responsible for the final content, claims, and outcomes of this article.

REFERENCES

Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Min Wu, Ming-Ming Cheng, Ender Konukoglu, and Serge Belongie. Multimodality helps few-shot 3d point cloud semantic segmentation. In *International Conference on Learning Representations*, volume 2025, pp. 55414–55435, 2025.

Petra Bevandić and Siniša Šegvić. Automatic universal taxonomies for multi-domain semantic segmentation. *arXiv preprint arXiv:2207.08445*, 2022.

Petra Bevandić, Marin Oršić, Ivan Grubišić, Josip Šarić, and Siniša Šegvić. Multi-domain semantic segmentation with overlapping labels. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2422–2431. IEEE, 2022.

Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *Journal of Visual Communication and Image Representation*, 109:104429, 2025.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

Linwei Chen, Ying Fu, Lin Gu, Dezhi Zheng, and Jifeng Dai. Spatial frequency modulation for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Xuweiyi Chen, Wentao Zhou, Aruni RoyChowdhury, and Zezhou Cheng. Point-moe: Large-scale multi-dataset training with mixture-of-experts for 3d semantic segmentation. In *International Conference on Learning Representations*, volume 2026, pp. 28106–28122, 2026.

Jiho Choi, Seonho Lee, Minhyun Lee, Seungho Lee, and Hyunjung Shim. Fine-grained image-text correspondence with cost aggregation for open-vocabulary part segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9782–9793. IEEE, 2025.

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*, pp. 3213–3223, 2016.

Junjie Fu, Chenliang Wang, Mingzhe Liu, Xinghua Li, Yi Liu, Wenjiao Shi, and Ruili Wang. Hyperr3snet: Leveraging hyperbolic space and vision foundation models for remote sensing semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

Feng Gao, Miao Fu, Jingchao Cao, Junyu Dong, and Qian Du. Adaptive frequency enhancement network for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025.

Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5108–5115. IEEE, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Lukas Hoyer, David Joseph Tan, Muhammad Ferjad Naeem, Luc Van Gool, and Federico Tombari. Semivl: semi-supervised semantic segmentation with vision-language guidance. In *European Conference on Computer Vision*, pp. 257–275. Springer, 2024.

Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 1769–1776. IEEE, 2020.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. Is clip ideal? no. can we fix it? yes! In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22436–22446, 2025.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36:29914–29934, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: A composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2879–2888, 2020.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

Jindong Li, Yongguang Li, Yali Fu, Jiahong Liu, Yixin Liu, Menglin Yang, and Irwin King. Clip-powered domain generalization and domain adaptation: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27948–27959. IEEE, 2024.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 5792–5801. IEEE, 2022.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Rong Ma, Jie Chen, Xiangyang Xue, and Jian Pu. Automated label unification for multi-dataset semantic segmentation with gnns. *Advances in Neural Information Processing Systems*, 37: 80141–80164, 2024.

Ishan Nigam, Chen Huang, and Deva Ramanan. Ensemble knowledge transfer for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1499–1508. IEEE, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Han Sun, Rui Gong, Ismail Nejjar, and Olga Fink. Dynalign: Unsupervised dynamic taxonomy alignment for cross-domain segmentation. In *International Conference on Learning Representations*, volume 2025, pp. 78191–78214, 2025.

Qiang Wan, Zilong Huang, Jiachen Lu, Gang Yu, and Li Zhang. Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation. In *The eleventh international conference on learning representations*, 2023.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

Tianyi Wu, Sheng Tang, Rui Zhang, Juan Cao, and Yongdong Zhang. Cgnet: A light-weight context guided network for semantic segmentation. *IEEE Transactions on Image Processing*, 30: 1169–1179, 2020.

Chenxi Xie, Minghan Li, Hui Zeng, Jun Luo, and Lei Zhang. Mass13k: a matting-level semantic segmentation benchmark. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14046–14056. IEEE, 2025.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5987–5995. IEEE, 2017.

Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *Visual Intelligence*, 4(1):2, 2026.

Guoyu Yang, Yuan Wang, Daming Shi, and Yanzhong Wang. Golden cudgel network for real-time semantic segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25367–25376. IEEE, 2025.

Bingfeng Zhang, Siyue Yu, Jimin Xiao, Yunchao Wei, and Yao Zhao. Frozen clip-dino: A strong backbone for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):4198–4214, 2025a.

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023a.

Dengke Zhang, Fagui Liu, and Quan Tang. Corrclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24677–24687. IEEE, 2025b.

Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2735–2745. IEEE, 2022a.

Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1020–1031. IEEE, 2023b.

Shi-Chen Zhang, Yunheng Li, Yu-Huan Wu, Qibin Hou, and Ming-Ming Cheng. Revisiting efficient semantic segmentation: Learning offsets for better spatial and class feature alignment. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22361–22371. IEEE, 2025c.

Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. Topformer: Token pyramid transformer for mobile semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12073–12083. IEEE, 2022b.

Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *European conference on computer vision*, pp. 418–434. Springer, 2018.

Qiang Zhou, Yuang Liu, Chaohui Yu, Jingliang Li, Zhibin Wang, and Fan Wang. Lmseg: Language-guided multi-dataset segmentation. *arXiv preprint arXiv:2302.13495*, 2023.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15116–15127. IEEE, 2023.

A Semantic Atom Vocabulary

The 39 atoms, in checkpoint order, are:

> unknown, person, rider, bicycle, motorcycle, car, truck, bus, train, vehicle-other, drone, boat, animal, road, sidewalk, curve, car-stop, guardrail, color-cone, bump, obstacle, construction, building, wall, fence, pole, traffic-light, traffic-sign, vegetation, terrain, sky, water, human-diver, aquatic-plant/reef, wreck/ruin, robot/instrument, reef-invertebrate, fish-vertebrate, sand/rock.

B Representative Class–Atom Relations

Table 4: Examples from the sparse initialization. These are trainable preferences, not hard rules. The region-conditioned term can change their relative contribution spatially.

Native label	Compatible atoms	Ambiguity
VDD vehicle	car, truck, bus, train, bicycle, other	coarse merge
MFNet car	car, truck, bus, vehicle-other	coarse merge
MFNet unlabeled	unknown and scene-content atoms	policy-dependent
AeroScapes construction	construction, building, wall, fence, pole	structural merge
FMB rider	rider, bicycle, motorcycle	actor/vehicle merge
SUIM wreck/ruin	wreck/ruin, construction	domain-specific structure

C Detailed Derivation of Reversible Taxonomy Transport

This section expands the forward model, its optimization signal, and the limits of the reverse constraint. The derivation follows the implemented computation exactly. Let $C_d = |\mathcal{C}_d|$, $K = 39$, $S = 8$, and let Ω be the one-eighth-resolution decoder lattice. The indices c, z, u, and r select a native class, atom, low-rank factor, and lattice site. The notation $|\Omega|$ is the number of sites. Vectors are column vectors, and $\top$ denotes transpose. The operator $\mathrm{Diag}(v)$ places the entries of v on a diagonal. The set Δ^{n-1} is the probability simplex over n outcomes. Every matrix $T_{d,r} \in \mathbb{R}^{C_d \times K}$ is column-stochastic. Thus, its entries are nonnegative and each column sums to one.

C.1 Multiscale atom posterior

For MiT stages $i \in \{1, 2, 3, 4\}$, write the encoder tensor as F_i and its learned 1×1 projection as P_i. Bilinear resizing $\mathcal{R}_i$ aligns every projected tensor to the second-stage lattice. The fused feature and atom posterior are

$$h = \psi(\mathcal{R}_1 P_1(F_1) \parallel P_2(F_2) \parallel \mathcal{R}_3 P_3(F_3) \parallel \mathcal{R}_4 P_4(F_4)), \tag{14}$$

$$\pi_{r,z} = p_\theta(z \mid x, r) = \frac{\exp(w_z^\top h_r + b_z)}{\sum_{j=1}^{K} \exp(w_j^\top h_r + b_j)}, \qquad r \in \Omega, \tag{15}$$

Here $\parallel$ denotes channel concatenation. The tensor h is the fused 128-channel feature. The block ψ contains the implemented 3×3 convolution, group normalization, GELU, and spatial dropout. The vector $w_z \in \mathbb{R}^{128}$ and scalar b_z are the atom classifier parameters. The index j is a dummy atom index in the softmax denominator. This establishes a normalized latent distribution before any native class is introduced. Consequently, the same 39 coordinates retain a common ontology index across dataset-specific checkpoints even though their learned visual parameters are independent.

C.2 Reliability-gated geometry

Let $s_r \in \mathbb{R}^{128}$ be the projected output of the frozen MobileSAM image encoder at r. The gate network g_ρ maps the concatenated visual and SAM vectors to one scalar. The matrices $W_v, W_s \in \mathbb{R}^{S \times 128}$ and biases $b_v, b_s \in \mathbb{R}^S$ form the two region projections. The region state is

$$\begin{aligned} \rho_r &= \sigma(g_\rho(h_r \parallel s_r)), \\ q_r &= \tanh(W_v h_r + b_v + \rho_r (W_s s_r + b_s)). \end{aligned} \tag{16}$$

The logistic sigmoid σ makes the reliability gate $\rho_r \in (0,1)$. The vector $q_r \in [-1,1]^S$ is the region state. The gate controls only the SAM contribution. Visual evidence $W_v h_r + b_v$ is never removed. The outer $\tanh$ bounds every component of q_r. It therefore bounds the dynamic logit correction $\Delta_{d,r,c,z}$ by

$$|\Delta_{d,r,c,z}| = \left| \sum_{u=1}^{S} q_{r,u} U_{d,c,u} V_{z,u} \right| \leq \sum_{u=1}^{S} |U_{d,c,u} V_{z,u}|. \tag{17}$$

The function $\tanh$ is the hyperbolic tangent. Vertical bars around a scalar denote absolute value. MobileSAM is evaluated under no-gradient mode. Nevertheless, ρ_r and the SAM projection receive supervision through the native segmentation, cycle, and spatial losses. The gate is therefore an optimized trust variable. It is not a confidence score imported from SAM. It is also not an independently supervised mask.

C.3 Text calibration and auditable initialization

Let e_c and e_z be unit-normalized frozen CLIP text embeddings of a native class description and an atom description. The offline cosine scores are standardized separately for each native class:

$$\kappa_{c,z} = e_c^\top e_z, \qquad \mu_c = \frac{1}{K} \sum_{j=1}^{K} \kappa_{c,j},$$

$$(\nu_c)^2 = \frac{1}{K} \sum_{j=1}^{K} (\kappa_{c,j} - \mu_c)^2, \qquad S^{\text{text}}_{d,c,z} = \frac{\kappa_{c,z} - \mu_c}{\nu_c + 10^{-6}}. \tag{18}$$

The score $\kappa_{c,z}$ is cosine affinity. The terms μ_c and ν_c are its mean and standard deviation over atoms for class c. The constant 10^{-6} prevents division by zero. The symbol j is a dummy atom index. Per-class standardization removes arbitrary score offsets and scales. Thus, the single learned coefficient $\alpha = \mathrm{softplus}(\tilde{\alpha})$ has a comparable role across native labels. At initialization, the auditable matrix A_d assigns $+2.5$ to specified compatible relations and -2.5 to the others. Before text and dynamic corrections, this is a log-odds gap of 5, or an odds ratio of $\exp(5) \approx 148.4$ between an otherwise matched compatible and incompatible native class for the same atom. These values are initial conditions rather than hard constraints. The matrix A_d, scalar α, and dynamic factors all remain trainable. The unconstrained scalar $\tilde{\alpha}$ is converted by $\mathrm{softplus}(t) = \log(1 + \exp t)$ for a scalar t. Thus, α is nonnegative.

C.4 Rank-controlled local transport

At each r, the unnormalized compiler is

$$L_{d,r} = A_d + \alpha S^{\text{text}}_d + \underbrace{U_d \,\mathrm{Diag}(q_r) V^\top}_{\Delta_{d,r}}, \qquad T_{d,r}(:,z) = \mathrm{softmax}\big(L_{d,r}(:,z)\big). \tag{19}$$

The matrix $L_{d,r}$ contains unnormalized class–atom logits. Its dynamic part is $\Delta_{d,r}$. The colon in $T_{d,r}(:,z)$ selects every native class in column z. The softmax normalizes that selected column. The function rank counts linearly independent matrix directions. Because $U_d \in \mathbb{R}^{C_d \times S}$ and $V \in \mathbb{R}^{K \times S}$,

$$\mathrm{rank}(\Delta_{d,r}) \leq \min\{\mathrm{rank}(U_d), \mathrm{rank}(\mathrm{Diag}(q_r)), \mathrm{rank}(V)\} \leq S. \tag{20}$$

Thus the static component can represent an unrestricted dataset taxonomy, whereas every spatial departure is assembled from at most eight shared class–atom interaction directions. The dynamic factors require $S(C_d + K)$ parameters rather than predicting an unconstrained $C_d K$ table from every pixel. For two locations r and r', their difference is

$$L_{d,r} - L_{d,r'} = U_d \,\mathrm{Diag}(q_r - q_{r'}) V^\top, \tag{21}$$

which makes clear that spatial variation is caused only by image-conditioned region coordinates. Neither the sparse initialization nor CLIP can vary with the image.

C.5 NATIVE SUPERVISION AS LATENT RESPONSIBILITY

For a valid pixel, let y_r be its target class and let $\pi_{r,z}$ be its atom posterior. The scalar m_r is the probability assigned to the target: $m_r = p(y_r \mid x, r) = \sum_z T_{d,r}(y_r \mid z)\pi_{r,z}$. Define

$$\gamma_{r,z} = p(z \mid y_r, x, r) = \frac{T_{d,r}(y_r \mid z)\pi_{r,z}}{m_r}, \qquad \sum_z \gamma_{r,z} = 1. \tag{22}$$

The quantity $\gamma_{r,z}$ is the exact posterior responsibility of atom z. The sums over z cover all K atoms. Define the atom logit as $a_{r,z}$ and the transport logit as $\ell_{r,c,z}$. Differentiating the per-pixel loss $\mathcal{L}_r = -\log m_r$ through both softmaxes yields

$$\frac{\partial \mathcal{L}_r}{\partial a_{r,z}} = \pi_{r,z} - \gamma_{r,z}, \tag{23}$$

$$\frac{\partial \mathcal{L}_r}{\partial \ell_{r,c,z}} = \gamma_{r,z}\left(T_{d,r}(c \mid z) - \mathbb{1}[c = y_r]\right). \tag{24}$$

The indicator $\mathbb{1}[c = y_r]$ is one when the condition holds and zero otherwise. Equation 23 moves the visual posterior toward the atoms that can explain the observed native label. Equation 24 updates a transport column in proportion to that atom's responsibility. The gradient reaching a region coordinate is therefore

$$\frac{\partial \mathcal{L}_r}{\partial q_{r,u}} = \sum_{c=1}^{C_d}\sum_{z=1}^{K} \frac{\partial \mathcal{L}_r}{\partial \ell_{r,c,z}} U_{d,c,u} V_{z,u}. \tag{25}$$

The symbol ∂ denotes a partial derivative. Similarly, $\partial \mathcal{L}_r / \partial \tilde{\alpha} = \sigma(\tilde{\alpha}) \sum_{c,z} (\partial \mathcal{L}_r / \partial \ell_{r,c,z}) S^{\text{text}}_{d,c,z}$, where the double sum covers all native classes and atoms. The logistic sigmoid $\sigma(\tilde{\alpha})$ is the derivative of softplus. The model can consequently suppress CLIP when its fixed affinities conflict with supervised pixels. Text supplies an initialization direction, but it never supplies a pixel label or an independent prediction path.

C.6 CYCLE CONSISTENCY AS NORMALIZED GRAM TRANSPORT

Fix one dataset and one location, then omit d and r. Let $t_z = T(:, z) \in \Delta^{C_d - 1}$ be transport column z. Let π be the atom posterior and p be the native-class posterior. Forward transport and the implemented reverse step form

$$p = T\pi, \qquad \widehat{\pi} = \mathcal{N}(T^\top p) = \mathcal{N}(G\pi), \qquad G = T^\top T, \quad G_{z,z'} = \langle t_z, t_{z'} \rangle, \tag{26}$$

Here $\widehat{\pi}$ is the reconstructed atom posterior. The matrix G is the Gram matrix of the transport columns, and $\langle \cdot, \cdot \rangle$ is the Euclidean inner product. The normalization is $\mathcal{N}(v) = v/(\mathbf{1}^\top v)$, where $\mathbf{1}$ is an all-ones vector. The Gram matrix exposes a failure mode hidden by forward supervision. If two atoms have identical compiler columns, their rows and columns in G coincide. The reverse path then cannot distinguish their mass. Minimizing

$$\mathcal{L}_{\text{cycle}} = \frac{1}{|\Omega|} \sum_{r \in \Omega} D_{\text{KL}}(\text{stopgrad}(\pi_r) \,\|\, \mathcal{N}(G_r \pi_r)) \tag{27}$$

therefore discourages overlapping columns when the corresponding atoms carry different mass on real images. The divergence D_{KL} is computed over atom components. In particular, $D_{\text{KL}}(P \| Q) = \sum_z P_z \log(P_z/Q_z)$ for atom distributions P and Q. The operator stopgrad preserves its input value but blocks its gradient. This fixed target prevents the atom predictor from moving merely to match a weak reverse path.

This constraint is deliberately local rather than globally invertible. When $C_d < K$, $\text{rank}(G_r) \leq C_d < K$, so G_r cannot equal the $K \times K$ identity over the whole simplex. Moreover, stochastic transport obeys the data-processing inequality. For any two atom distributions π and π', it gives

$$D_{\text{KL}}(T_r \pi \,\|\, T_r \pi') \leq D_{\text{KL}}(\pi \,\|\, \pi'), \tag{28}$$

and cannot preserve arbitrary distinctions discarded by a coarse taxonomy. The attainable objective is approximate recovery on the data-supported atom manifold. For each pixel, Pinsker's inequality also connects the optimized KL term to the displayed diagnostic residual:

$$\tfrac{1}{2}\|\pi_r - \widehat{\pi}_r\|_1 \leq \sqrt{\tfrac{1}{2} D_{\text{KL}}(\pi_r \,\|\, \widehat{\pi}_r)}. \tag{29}$$

The norm $\|v\|_1 = \sum_z |v_z|$ is the ℓ_1 norm. The square root is the nonnegative principal root.

C.7 Spatial and boundary regularization

For a multichannel field $F \in \mathbb{R}^{J\times H'\times W'}$, J is the channel count and $H' \times W'$ is the decoder resolution. Let $D_x F$ and $D_y F$ be forward horizontal and vertical differences with zero padding at the missing predecessor. The implementation uses

$$\mathcal{L}_{\text{region-TV}} = \operatorname{mean}|D_x q| + \operatorname{mean}|D_y q|, \tag{30}$$

The operator mean averages every valid field element, and $|\cdot|$ is elementwise absolute value. We then form a soft boundary magnitude

$$b_r(F) = \operatorname{clip}_{[\varepsilon,1-\varepsilon]}\left(\frac{1}{J}\sum_{j=1}^{J}\big(|D_x F_{j,r}| + |D_y F_{j,r}|\big)\right), \qquad \varepsilon = 10^{-5}. \tag{31}$$

Here j indexes channels. The clipping operator restricts its argument to the displayed interval, and ε is its numerical margin. The target boundary b_r^* is one when either valid horizontal or vertical neighbor has a different native label, and zero otherwise. With BCE denoting binary cross entropy,

$$\begin{aligned}\mathcal{L}_{\text{boundary}} &= \operatorname{BCE}(b(p_d), b^*),\\ \mathcal{L}_{\text{region-boundary}} &= \operatorname{BCE}(b(q), b^*).\end{aligned} \tag{32}$$

Here p_d is the native probability field. The expressions $b(p_d)$ and $b(q)$ apply the boundary operator at every lattice site. The function $\operatorname{BCE}(u, v)$ is mean binary cross entropy between a predicted field u and target field v. The first loss aligns the final native probability transition with the mask. The second makes changes in the region coordinates coincide with places where the annotation policy changes class. Total variation prevents unstructured pixel-wise oscillation, while the boundary term prevents excessive smoothing across genuine class transitions.

The remaining entropy penalty is the mean conditional entropy of every transport column,

$$\mathcal{L}_{\text{transport-ent}} = -\frac{1}{|\Omega|K}\sum_{r\in\Omega}\sum_{z=1}^{K}\sum_{c=1}^{C_d} T_{d,r}(c \mid z)\log T_{d,r}(c \mid z). \tag{33}$$

The logarithm is natural, and all indices in the displayed sums are shown explicitly. The loss favors decisive atom-to-class routing but does not determine which class is selected. Native likelihood, the initialized relation, and the cycle constraint provide that semantic direction.

C.8 Optimization and inference path

One training step is fully differentiable after the two frozen priors are loaded. First, MiT features are fused into h and π. Next, the frozen MobileSAM feature is projected and gated into q. Equation 19 then materializes the local transport. Native masks supervise $p_d = T_d\pi$. Finally, the five auxiliary terms in Equation 13 update the visual decoder, gate, base relation, text scale, and low-rank factors. MobileSAM weights and the stored CLIP affinity receive no gradients.

For a decoder lattice of $N = H'W'$ locations, the explicit dynamic contraction cost is $O(NC_dKS)$ and the transport tensor occupies $O(NC_dK)$ values. The notation $O(\cdot)$ denotes asymptotic order. Here $K = 39$, $S = 8$, and $C_d \le 19$. At inference, the cycle and all loss-only operators are removed. The model computes the image features, frozen SAM geometry, atom posterior, and forward transport once, upsamples $\log p_d$, the elementwise log posterior, to the input resolution. It then returns the native class with the largest probability. CLIP is not executed online because its class–atom affinity is already stored in the checkpoint input bank.

D Additional Overall Metrics

Tables 5 and 6 report the four complete-split metrics omitted from the main comparison for legibility. Their splits, predictions, and class orderings are identical to those used for Table 1.

Table 5: Overall accuracy (aAcc), macro recall (mAcc), macro precision (mPrec), and frequency-weighted IoU (fwIoU) on AeroScapes, VDD, and MFNet (%). Splits and evaluation predictions match Table 1. Column bests are bold and second-best values are underlined.

Method	AeroScapes				VDD				MFNet			
	aAcc	mAcc	mPrec	fwIoU	aAcc	mAcc	mPrec	fwIoU	aAcc	mAcc	mPrec	fwIoU
CGNet	87.57	49.49	51.48	79.34	80.32	68.86	73.52	68.06	95.60	50.89	62.89	92.25
HRNet	92.19	77.65	84.36	85.66	92.74	89.55	89.87	86.78	98.09	79.59	84.96	96.45
ICNet	95.17	75.98	78.24	91.04	89.27	81.46	83.67	81.39	95.36	64.21	70.62	92.35
SegFormer-B3	95.79	73.98	82.19	92.14	88.74	79.99	82.75	80.59	97.23	67.94	81.62	94.90
SegFormer-B4	95.55	72.08	79.15	91.73	89.15	81.09	82.68	81.26	97.20	68.22	80.80	94.87
R101-SegFormer	88.94	58.69	64.45	80.59	87.68	79.85	84.29	78.85	96.29	61.67	73.91	93.54
R50-DLV2	88.11	66.39	63.67	79.86	87.29	77.81	82.19	78.21	96.43	61.42	74.61	93.74
ResNeSt-50-DLV2	94.77	81.15	85.55	90.24	92.07	87.59	89.88	85.76	98.06	75.19	**87.39**	96.34
ResNeSt-101-DLV2	96.13	84.04	85.61	92.68	92.69	89.05	90.17	86.75	98.16	76.93	87.28	96.53
ResNeSt-200-DLV2	96.26	87.19	88.52	92.88	92.87	89.74	90.41	87.00	98.21	78.00	87.24	96.63
SeaFormer	82.51	38.78	56.01	71.38	76.22	59.82	64.60	62.66	95.16	34.71	56.41	91.18
TopFormer	79.36	33.78	61.78	66.77	82.10	68.90	72.30	70.78	95.50	40.37	68.52	91.81
ResNeXt-50-DLV2	94.55	84.17	86.38	89.77	92.21	88.43	89.38	85.95	97.76	79.40	82.59	95.87
ResNeXt-101-DLV2	94.35	81.93	84.20	89.51	91.97	89.03	88.31	85.59	98.01	77.79	84.74	96.30
MaSSFormer	93.48	84.31	80.58	88.20	86.83	84.21	87.05	78.14	97.56	72.45	71.00	95.72
SAM2-UNet	95.60	80.73	88.92	91.70	91.87	86.64	89.17	85.36	98.14	80.41	82.65	96.59
HQ-SAM	91.17	71.04	77.57	84.46	88.58	78.92	85.67	80.52	97.18	59.43	78.99	94.75
LSeg	96.16	75.32	83.09	92.76	90.40	80.28	84.14	83.08	97.37	76.31	79.35	95.22
SemiVL	95.31	76.77	83.22	91.22	93.01	90.68	89.77	87.22	98.03	77.19	84.80	96.33
PartCATSeg	94.13	77.04	79.27	89.15	85.81	74.57	77.55	75.75	97.29	77.98	77.95	95.18
ReTax-Seg	97.10	87.51	89.22	94.44	93.96	90.97	91.76	88.84	98.22	81.18	85.00	96.71
ReTax-Seg-C	**97.62**	**88.38**	**89.71**	**94.92**	**94.08**	**91.26**	**92.08**	**89.36**	**98.49**	**81.74**	86.77	**96.82**

Table 6: Overall accuracy (aAcc), macro recall (mAcc), macro precision (mPrec), and frequency-weighted IoU (fwIoU) on FMB, Cityscapes, and SUIM (%). Splits and evaluation predictions match Table 1. Column bests are bold and second-best values are underlined.

Method	FMB				Cityscapes				SUIM			
	aAcc	mAcc	mPrec	fwIoU	aAcc	mAcc	mPrec	fwIoU	aAcc	mAcc	mPrec	fwIoU
CGNet	84.62	45.42	53.81	74.33	82.99	42.63	47.11	73.69	67.97	46.18	50.03	54.39
HRNet	89.81	61.28	70.96	82.09	91.06	69.01	69.17	84.97	80.62	58.66	70.28	67.93
ICNet	85.89	48.64	58.37	76.06	84.86	46.18	52.26	75.81	66.31	43.91	54.98	52.77
SegFormer-B3	87.52	48.94	59.16	78.61	89.15	59.21	62.38	82.20	75.49	56.93	71.39	62.79
SegFormer-B4	87.71	49.68	60.26	78.89	89.06	59.40	60.67	82.22	76.42	58.31	72.89	63.97
R101-SegFormer	86.10	50.94	60.98	76.34	89.10	57.44	60.45	81.97	66.19	42.45	48.55	52.71
R50-DLV2	85.01	47.60	64.34	74.79	88.62	61.63	61.27	81.23	76.52	61.21	66.94	62.56
ResNeSt-50-DLV2	89.27	55.97	74.20	81.22	89.18	68.17	73.82	81.48	76.98	62.04	67.94	63.48
ResNeSt-101-DLV2	89.72	57.61	81.02	81.83	90.84	74.04	73.31	84.28	79.20	65.43	75.93	66.03
ResNeSt-200-DLV2	89.99	58.22	80.71	82.27	90.38	72.62	72.71	83.54	80.45	66.90	74.88	67.69
SeaFormer	65.47	25.68	54.05	49.15	82.57	38.47	46.52	72.77	70.22	44.73	58.21	55.70
TopFormer	82.15	39.68	50.05	70.70	82.19	36.56	44.22	71.51	69.60	43.67	60.98	54.94
ResNeXt-50-DLV2	89.50	58.99	79.08	81.52	91.68	74.44	74.96	85.60	79.52	62.06	70.92	66.48
ResNeXt-101-DLV2	90.20	59.77	80.47	82.65	92.19	74.27	77.61	86.40	78.80	63.26	68.57	65.49
MaSSFormer	89.65	63.62	71.85	82.01	62.93	27.18	27.38	58.19	78.53	59.60	67.61	66.28
SAM2-UNet	90.99	68.07	**81.61**	83.90	93.57	75.08	80.85	88.47	87.22	79.33	82.50	77.73
HQ-SAM	87.46	51.79	74.13	78.24	90.39	60.49	70.59	83.32	80.18	64.68	68.85	67.85
LSeg	89.11	60.77	75.75	80.99	91.57	65.44	74.99	85.25	28.68	35.29	40.76	21.62
SemiVL	89.55	59.18	76.14	81.62	94.87	80.75	86.04	90.61	86.98	77.18	86.43	77.24
PartCATSeg	86.11	58.88	68.01	77.13	93.33	75.03	79.64	88.24	86.68	78.45	84.63	76.71
ReTax-Seg	91.34	68.74	80.31	84.51	95.01	80.46	86.70	90.83	86.96	81.36	86.26	77.37
ReTax-Seg-C	**91.72**	**69.84**	81.49	**84.93**	**95.71**	**81.16**	**87.04**	**91.25**	**87.49**	**81.92**	**86.83**	**78.14**

Table 7: Zero-shot transfer mIoU (%). Each target averages five foreign-source checkpoints. No target tuning is used.

Method	Aero.	VDD	MFNet	FMB	City.	SUIM	Mean
CGNet	5.60	6.88	10.42	1.52	4.50	6.30	5.87
HRNet	8.25	4.80	14.60	2.55	7.84	6.59	7.44
ICNet	6.52	6.90	11.31	2.02	4.23	5.50	6.08
SegFormer-B3	6.58	6.33	13.25	2.36	6.14	7.04	6.95
SegFormer-B4	6.76	6.95	13.03	2.39	5.94	6.87	6.99
R101-SegFormer	6.22	7.49	12.35	2.10	4.45	6.01	6.44
ResNeSt-200-DLV2	8.20	5.86	14.94	3.01	7.24	7.12	7.73
HQ-SAM	10.86	7.69	**16.10**	3.23	**8.45**	8.11	9.07
ReTax-Seg	**11.83**	**9.52**	15.29	**3.94**	6.64	**9.50**	**9.45**

E Additional Controlled Experiments

The following four controlled studies are placed immediately after the cross-dataset comparison because they test the semantic compiler itself rather than add another generic segmentation benchmark. All runs use the same MiT-B5 initialization, crop size, optimizer, schedule, checkpoint-selection rule, and random seed as the controlled ablation in Table 3.

E.1 Synthetic coarse–fine taxonomy conflict

We first reserve the held-out Cityscapes evaluation split. Only the training images are partitioned, with a fixed stratified split, into equal fine-policy and coarse-policy subsets. The fine subset retains the original 19 labels. In the coarse subset, car, truck, bus, motorcycle, and bicycle become one *vehicle* label; building, wall, fence, and pole become one *construction* label. The other labels are unchanged, giving 12 coarse classes. The original fine masks of coarse-policy training images are never used for optimization, probe fitting, or model selection. Each variant sees the same images and policy-specific supervision.

Table 8 separates two questions. *Coarse mIoU* evaluates the coarse-taxonomy output against the held-out fine masks after applying the same merges. To test whether the shared atom representation retains the distinctions hidden by the coarse policy, we freeze each trained model and extract its 39-dimensional pre-transport atom posterior. Two group-specific linear softmax probes are fitted only on fine-policy training pixels, using identical sampling and fitting settings for every variant. The five-way vehicle probe and four-way construction probe are then evaluated on held-out pixels whose original labels belong to the corresponding group. Each group score is the macro IoU over its fine classes; the last column averages the two group scores. Ground-truth group membership selects the probe and defines this diagnostic evaluation mask; the fine label is never fed to the probe. Thus these are oracle-group representation diagnostics, not end-to-end fine-label segmentation or discovery without fine supervision.

Table 8: Synthetic coarse–fine Cityscapes experiment (%). Coarse mIoU is computed over 12 classes on the held-out split. Vehicle and construction columns report fine-class macro IoU from frozen-atom linear probes, evaluated only within the respective ground-truth groups; their unweighted mean is shown last.

Variant	Coarse mIoU	Vehicle fine mIoU	Construction fine mIoU	Mean fine mIoU
Static class–atom mapping	78.02	17.76	37.49	27.62
Global learned transport	78.22	18.33	**37.59**	27.96
Dynamic transport, w/o cycle	**78.29**	18.51	37.48	28.00
Full dynamic ReTax-Seg	78.28	**18.55**	37.52	**28.03**

E.2 Does cycle consistency prevent atom collapse?

The segmentation ablation alone does not reveal whether removing the reverse constraint collapses the latent vocabulary. We therefore aggregate the complete evaluation split and report four direct utilization statistics. An atom is active when it is the top-1 atom for at least 0.1% of valid pixels. Usage perplexity is the exponentiated entropy of the marginal atom histogram. Effective rank is the entropy-based rank of the atom-posterior covariance matrix. An atom is dead when its mean posterior is below 10^{-4}. These fixed definitions prevent a favorable threshold from being chosen after observing the results.

Table 9: Cycle-consistency collapse audit. FMB and AeroScapes mIoU values and latent statistics are computed from matched runs over their complete evaluation splits. Higher is better except for dead-atom ratio.

Variant	FMB mIoU	Aero. mIoU	Active	Perplexity	Eff. rank	Dead (%)
Dynamic transport, w/o cycle	60.34	80.28	19	7.489	6.798	51.28
Full dynamic ReTax-Seg	**60.64**	**80.34**	19	**7.497**	**6.802**	51.28

E.3 Incorrect and perturbed CLIP priors

Because the CLIP affinity is an offline training prior rather than an online inference branch, each perturbation is retrained from scratch. The synonym condition replaces class names with manually verified equivalents, the shuffle condition permutes names within each native taxonomy, and the random condition replaces the normalized CLIP affinity with a norm-matched random matrix. This distinguishes robustness to wording from resistance to a semantically incorrect language prior.

Table 10: Training-time CLIP-prior perturbation on the two controlled-ablation datasets (%). Every row is retrained with the matched schedule and seed. α_{clip} is the learned global prior strength at the selected checkpoint. The correct-prior values are reproduced from Table 3.

Training prior	FMB mIoU	Aero. mIoU	Mean	α_{clip}
Correct class names	60.64	80.34	**70.49**	0.430
Verified synonyms	**60.66**	80.32	**70.49**	0.433
Within-taxonomy shuffled names	60.42	80.30	70.36	0.155
Norm-matched random affinity	60.35	**80.35**	70.35	0.218
No CLIP affinity	60.38	80.30	70.34	0

E.4 Atom-count and transport-rank sensitivity

We vary one capacity control at a time around the default configuration ($K = 39$ atoms and rank $S = 8$). Atom inventories with fewer than 39 entries are constructed by merging the least frequent compatible atoms; larger inventories split the most frequent ambiguous atoms. The same deterministic inventory is shared by all compared runs. Parameter count includes the complete model, and mean mIoU averages FMB and AeroScapes.

Table 11: One-factor-at-a-time sensitivity to semantic-atom count K and transport rank S. The default $K = 39, S = 8$ result is the same checkpoint as Table 3. All other settings retain the matched protocol.

Sweep	Setting	FMB mIoU	Aero. mIoU	Mean	Params. (M)
Atom count K ($S = 8$)	19	60.37	79.94	81.94	88.41
	29	60.48	80.31	82.31	88.41
	39	**60.64**	**80.34**	**70.49**	88.41
	59	60.42	80.22	70.32	88.41
Rank S ($K = 39$)	2	60.29	80.17	70.23	88.41
	4	60.35	80.16	70.26	88.41
	8	**60.64**	**80.34**	**70.49**	88.41
	16	60.59	80.30	70.45	88.41

F Result Provenance

Across Tables 1, 5, and 6, ReTax-Seg ranks first in 30 of the 36 dataset–metric comparisons. In Table 1, it achieves the highest mIoU on all six datasets. The values are 80.34, 84.50, 71.91, 60.64, 72.67, and 71.74% in dataset order. Their mean is 73.63%. Relative to the strongest competing result on each dataset, the corresponding margins are 1.28, 1.83, 1.22, 0.59, 0.17, and 2.32 points. ReTax-Seg also ranks first in mF1 on all six datasets. Its margins are 0.69, 1.17, 0.81, 0.18, 0.11, and 1.89 points, respectively.

For the two accuracy measures in Tables 5 and 6, ReTax-Seg leads aAcc on five datasets. They are AeroScapes, VDD, MFNet, FMB, and Cityscapes. It leads mAcc on AeroScapes, VDD, MFNet, FMB, and SUIM. Its only non-leading entries are Cityscapes mAcc, where 80.46% is 0.29 points below SemiVL, and SUIM aAcc, where 86.96% is 0.26 points below SAM2-UNet. The largest mAcc advantage is on SUIM: 81.36% versus 79.33% for the strongest alternative, a gain of 2.03 points. On MFNet, the aAcc lead is narrower at 98.22% versus 98.21%.

For mPrec and fwIoU, Tables 5 and 6 contain eight further wins. ReTax-Seg achieves the highest fwIoU on AeroScapes, VDD, MFNet, FMB, and Cityscapes. The values are 94.44, 88.84, 96.71, 84.51, and 90.83%. The gains over the next-best values are 1.56, 1.62, 0.08, 0.61, and 0.22 points. On SUIM, its 77.37% fwIoU is 0.36 points below SAM2-UNet. For mPrec, ReTax-Seg ranks first on AeroScapes (89.22%), VDD (91.76%), and Cityscapes (86.70%). It records 85.00% on MFNet, 80.31% on FMB, and 86.26% on SUIM, trailing the respective best values by 2.39, 1.30, and 0.17 points.

Table 12: Comparison of open-vocabulary segmentation methods (%). Baseline methods are MSeg (Lambert et al., 2020), NLL+ (Bevandić et al., 2022), X-Decoder (Zou et al., 2023), OpenSeeD (Zhang et al., 2023b), and OMG-Seg (Li et al., 2024).

Dataset	Method	aAcc	mAcc	mPrec	fwIoU	mF1	mIoU
AeroScapes	MSeg	80.09	50.26	75.53	68.81	55.63	43.74
	NLL+	96.00	88.15	**89.91**	94.26	87.37	80.58
	X-Decoder	97.52	86.23	89.31	93.22	85.03	79.42
	OpenSeeD	97.27	87.61	88.51	94.78	87.82	80.13
	OMG-Seg	97.19	87.79	88.58	94.59	87.62	78.16
	ReTax-Seg	97.10	87.51	89.22	94.44	88.28	80.34
	ReTax-Seg-C	**97.62**	**88.38**	89.71	**94.92**	**88.76**	**80.92**
VDD	MSeg	52.67	40.80	60.75	39.38	34.22	24.04
	NLL+	94.03	**91.50**	91.15	89.01	91.31	84.49
	X-Decoder	94.00	91.07	91.39	88.94	91.71	84.02
	OpenSeeD	93.46	91.23	91.49	88.64	91.30	84.19
	OMG-Seg	93.94	90.72	90.96	88.79	91.24	84.39
	ReTax-Seg	93.96	90.97	91.76	88.84	91.35	84.50
	ReTax-Seg-C	**94.08**	91.26	**92.08**	**89.36**	**92.03**	**84.97**
MFNet	MSeg	96.85	39.23	82.69	93.88	46.93	37.79
	NLL+	98.21	79.17	85.45	96.66	82.09	70.75
	X-Decoder	98.41	81.25	85.95	96.03	83.42	71.34
	OpenSeeD	98.29	80.77	83.59	95.90	82.01	70.75
	OMG-Seg	98.24	80.41	84.80	96.77	82.05	69.52
	ReTax-Seg	98.22	81.18	85.00	96.71	82.97	71.91
	ReTax-Seg-C	**98.49**	**81.74**	**86.77**	**96.82**	**83.55**	**72.38**
FMB	MSeg	22.31	22.64	47.95	19.37	17.92	12.46
	NLL+	91.32	68.24	78.90	84.58	72.22	60.36
	X-Decoder	91.46	67.41	80.64	84.69	71.42	61.21
	OpenSeeD	91.41	69.80	78.06	84.71	71.91	59.85
	OMG-Seg	91.71	69.43	72.64	82.21	66.82	55.71
	ReTax-Seg	91.34	68.74	80.31	84.51	72.69	60.64
	ReTax-Seg-C	**91.72**	69.84	**81.49**	**84.93**	**73.48**	**61.27**
Cityscapes	MSeg	90.76	47.79	75.11	83.49	50.62	41.54
	NLL+	94.82	77.54	86.39	90.49	81.23	70.00
	X-Decoder	94.01	79.83	84.22	82.63	80.93	71.86
	OpenSeeD	95.39	80.75	86.41	90.33	81.92	70.19
	OMG-Seg	95.38	78.20	85.29	90.25	82.87	69.34
	ReTax-Seg	95.01	80.46	86.70	90.83	83.18	72.67
	ReTax-Seg-C	**95.71**	**81.16**	**87.04**	**91.25**	**83.98**	**72.75**
SUIM	MSeg	40.92	18.30	65.89	18.09	16.63	10.29
	NLL+	87.27	81.44	85.91	77.95	81.53	71.35
	X-Decoder	86.59	80.27	83.70	77.18	82.65	71.53
	OpenSeeD	85.43	79.38	83.82	77.60	82.40	71.55
	OMG-Seg	86.17	80.59	84.17	75.60	81.22	70.36
	ReTax-Seg	86.96	81.36	86.26	77.37	82.30	71.74
	ReTax-Seg-C	**87.49**	**81.92**	**86.83**	**78.14**	**82.86**	**72.03**

Result analysis. ReTax-Seg-C ranks first in 34 of the 36 dataset–metric comparisons: all six metrics on MFNet, FMB, Cityscapes, and SUIM, and five metrics on AeroScapes and VDD. The exceptions are AeroScapes mPrec (89.71 versus 89.91 for NLL+) and VDD mAcc (91.26 versus 91.50 for NLL+). It also achieves the highest mIoU on all six datasets. Its average mIoU is 74.05%, improving over ReTax-Seg by 0.42 points (73.63%).

G Vision–Language Annotation and Representation Analysis

Figures 5 and 6 provide additional evidence for the semantic annotation process and the representation change induced by ReTax, respectively.



Figure 5: **Vision–language-assisted semantic annotation.** Representative visual concepts from heterogeneous domains are processed by Qwen3-VL-8B-Instruct to produce detailed image-level descriptions. The six examples below the pipeline illustrate the resulting annotations across aerial, urban, and underwater scenes. Purple dashed boxes identify the visual inputs associated with each description.

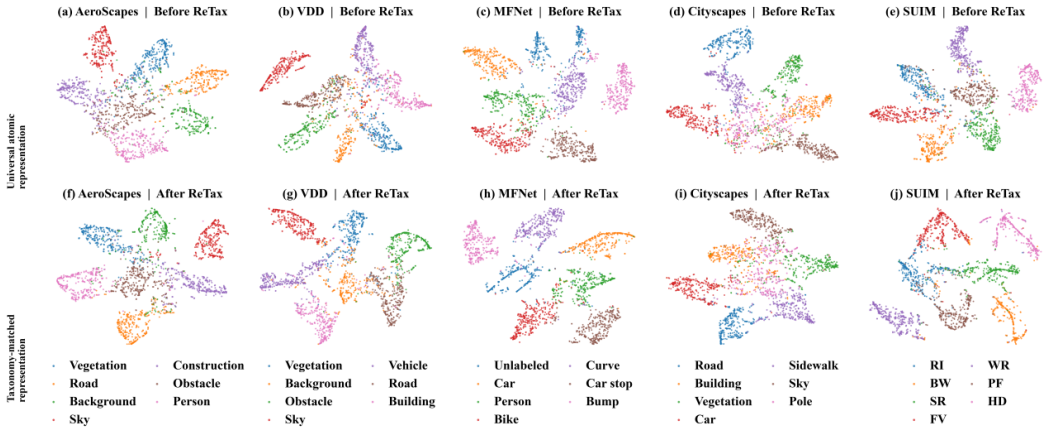

Figure 6: **t-SNE visualization before and after ReTax.** The top row shows the universal atomic representation before taxonomy adaptation, whereas the bottom row shows the corresponding taxonomy-matched representation after ReTax. Columns report AeroScapes, VDD, MFNet, Cityscapes, and SUIM. Colors denote the native semantic classes listed below each dataset. ReTax reorganizes the shared representation into more dataset-aligned class structures while retaining cross-domain semantic evidence.

H Extended Qualitative Evaluation

Figure 7 provides focused comparisons on VDD and FMB, while Figure 8 provides a detailed Cityscapes comparison. Figures 9, 10, and 11 add focused comparisons on AeroScapes, MFNet, and SUIM, respectively. Figures 14 and 16 extend the main-paper comparison with twelve disjoint images, two per dataset. Figures 13 and 15 provide enlarged split views of Figure 14. The first covers AeroScapes, VDD, and MFNet. The second covers FMB, Cityscapes, and SUIM. All panels use native dataset palettes. Colors are comparable within a dataset but not across datasets.

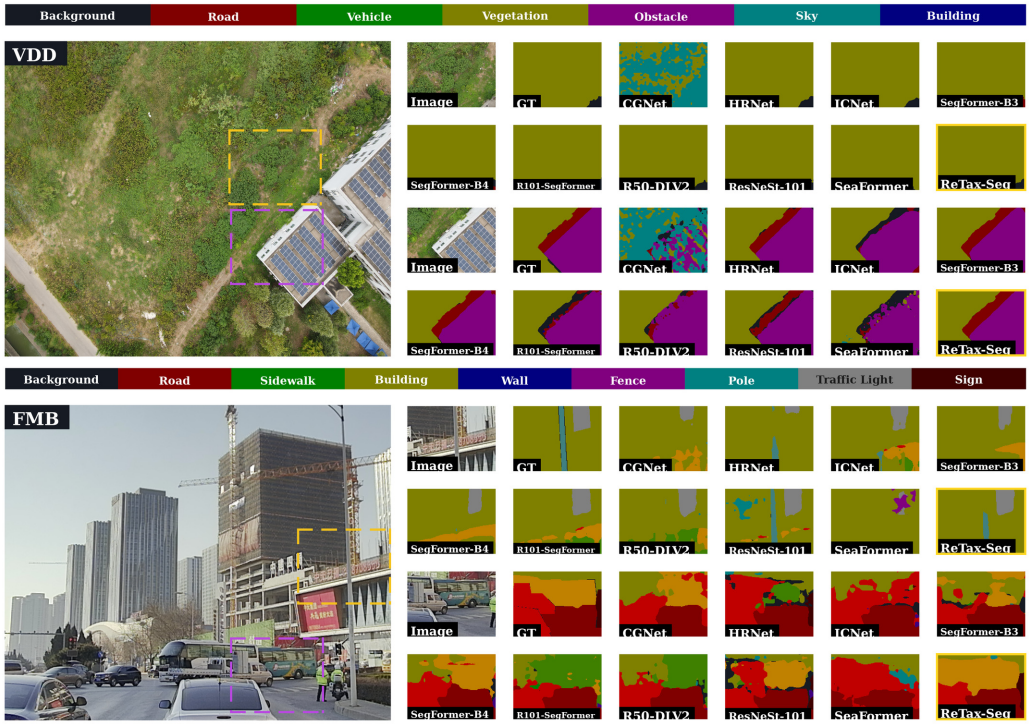

Figure 7: **Focused qualitative comparisons on VDD and FMB.** Each dataset panel shows the full image on the left and two enlarged regions on the right, linked by the magenta and yellow dashed boxes. For each region, the figure presents the RGB crop, native ground truth (GT), predictions from CGNet, HRNet, ICNet, SegFormer-B3, SegFormer-B4, R101-SegFormer, R50-DLV2, ResNeSt-101, SeaFormer, and ReTax-Seg. The color strips provide the class legend used by each panel, and yellow borders identify the ReTax-Seg predictions.

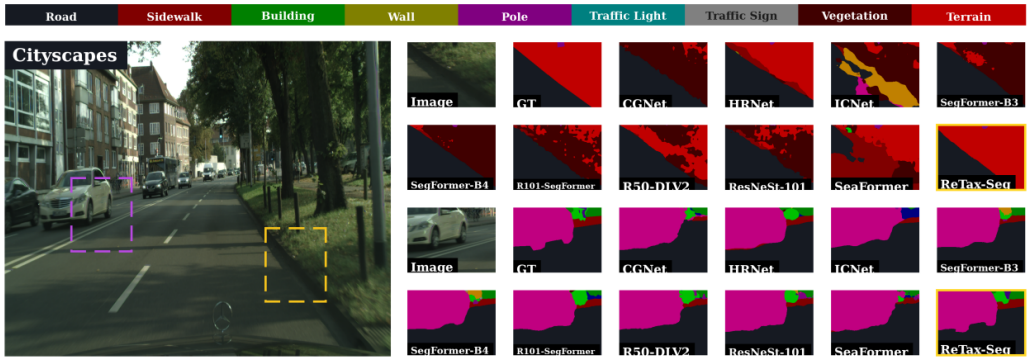

Figure 8: **Cityscapes comparison.** Magenta and yellow boxes mark two crops. Columns show RGB, ground truth, nine baselines, and ReTax-Seg. The top strip gives class colors. Yellow borders mark ReTax-Seg.



Figure 9: **Focused qualitative comparison on AeroScapes.** The left panel shows the full input image, with yellow and magenta dashed boxes marking two enlarged regions. For each region, the right panels show the RGB crop, native ground truth (GT), and predictions from CGNet, HRNet, ICNet, SegFormer-B3, SegFormer-B4, R101-SegFormer, R50-DLV2, ResNeSt-101, SeaFormer, and ReTax-Seg. The top strip gives the native class palette, and yellow borders identify the ReTax-Seg predictions.

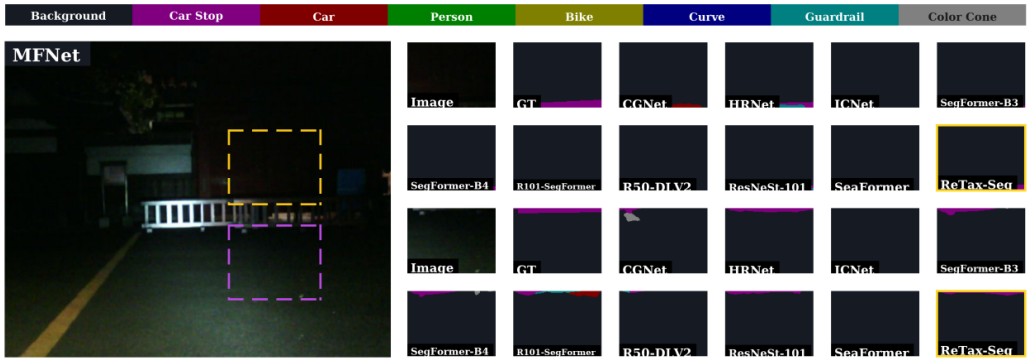

Figure 10: **Focused qualitative comparison on MFNet.** The layout and method order match Figure 9. Yellow and magenta dashed boxes mark two enlarged regions, the top strip gives the native class palette, and yellow borders identify the ReTax-Seg predictions.

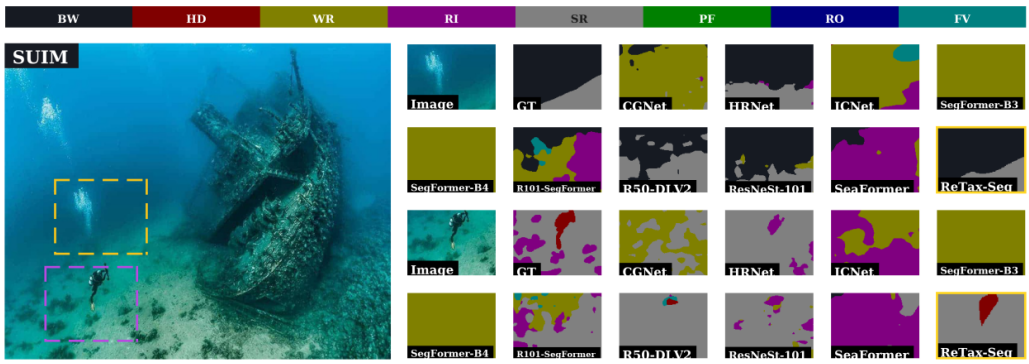

Figure 11: **Focused qualitative comparison on SUIM.** The layout and method order match Figure 9. Yellow and magenta dashed boxes mark two enlarged regions, the top strip gives the native class palette, and yellow borders identify the ReTax-Seg predictions.

Figure 12 provides a stage-wise view of feature propagation for AeroScapes, Cityscapes, and FMB. It follows the four encoder stages into the aligned decoder features and the final fused representation, then places the corresponding prediction beside the input and ground truth. This view complements the aggregate and qualitative results by showing how coarse semantic responses and finer spatial structure coexist before fusion in three distinct domains.

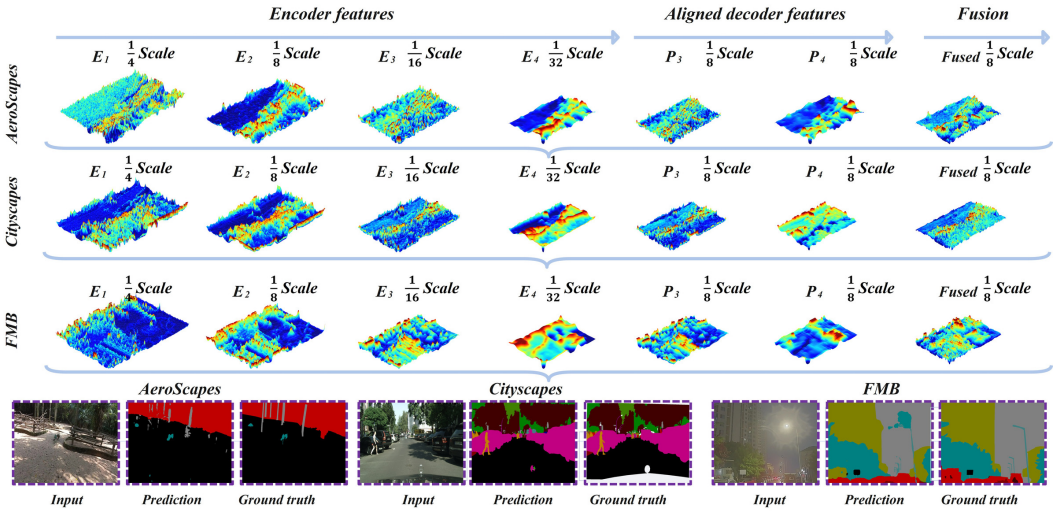

Figure 12: **Multiscale feature evolution and qualitative outputs.** Rows show AeroScapes, Cityscapes, and FMB. From left to right, the upper panels trace encoder features E_1–E_4 at $1/4$, $1/8$, $1/16$, and $1/32$ scale, aligned decoder features P_3 and P_4 at $1/8$ scale, and the fused $1/8$-scale representation. Surface height and color visualize relative activation magnitude within each panel and are not numerically calibrated across stages or datasets. The bottom triplets show the input, prediction, and ground truth for the same three datasets, linking the internal feature progression to the final segmentation output.

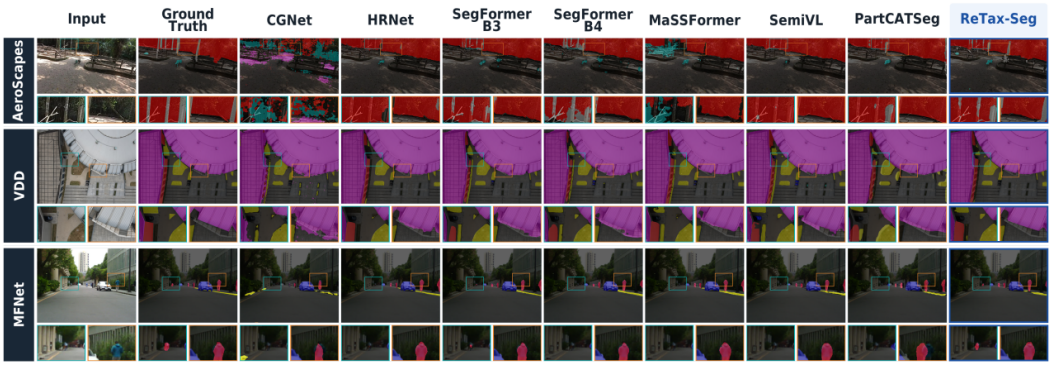

Figure 13: **Enlarged split view for AeroScapes, VDD, and MFNet.** This is the upper three-dataset portion of Figure 14. Columns show the input, ground truth, CGNet, HRNet, SegFormer-B3, SegFormer-B4, MaSSFormer, SemiVL, PartCATSeg, and ReTax-Seg. Each dataset block places the full prediction above two enlarged regions. Cyan and orange boxes identify the corresponding crops. The blue frame marks ReTax-Seg.

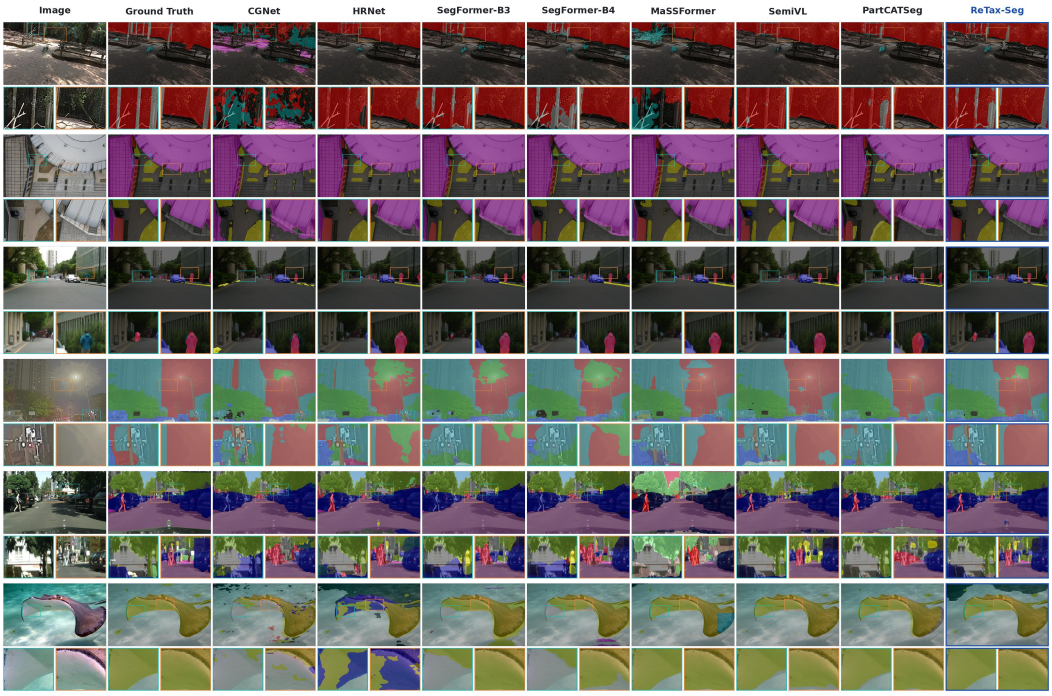

Figure 14: **Eight-method segmentation comparison with two enlarged details per image (set A).** Rows follow AeroScapes, VDD, MFNet, FMB, Cityscapes, and SUIM. The upper strip of each row is the full view. The lower strip shows the cyan and orange windows at a common scale. The blue frame marks ReTax-Seg and does not encode confidence.

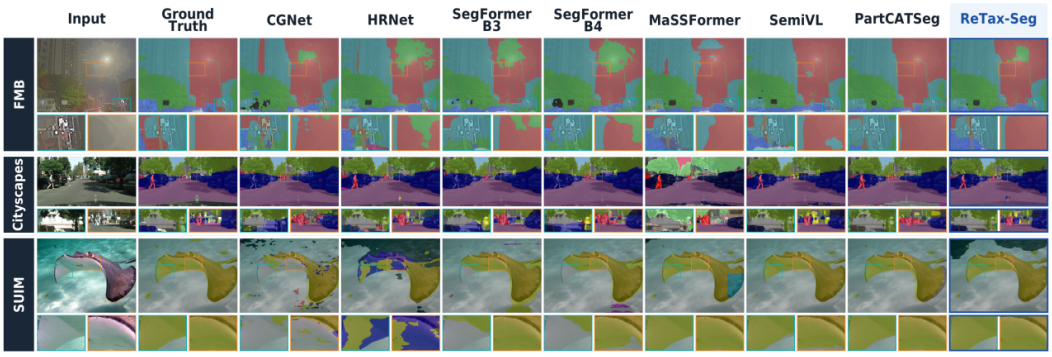

Figure 15: **Enlarged split view for FMB, Cityscapes, and SUIM.** This is the lower three-dataset portion of Figure 14. The methods and layout match Figure 13. Each dataset block contains the full prediction and two enlarged regions. Cyan and orange boxes identify the corresponding crops. The blue frame marks ReTax-Seg.

What the enlargements reveal. Across the twelve fixed images, the most visible differences occur at thin structures, object–background transitions, and boundaries between two semantically related classes. Several conventional models recover the dominant region but fragment a narrow object or extend a frequent class into an adjacent one. The text-conditioned baselines often assign a plausible semantic class while retaining a coarse spatial support. In contrast, the final column more often preserves a continuous small structure and follows the local image boundary.

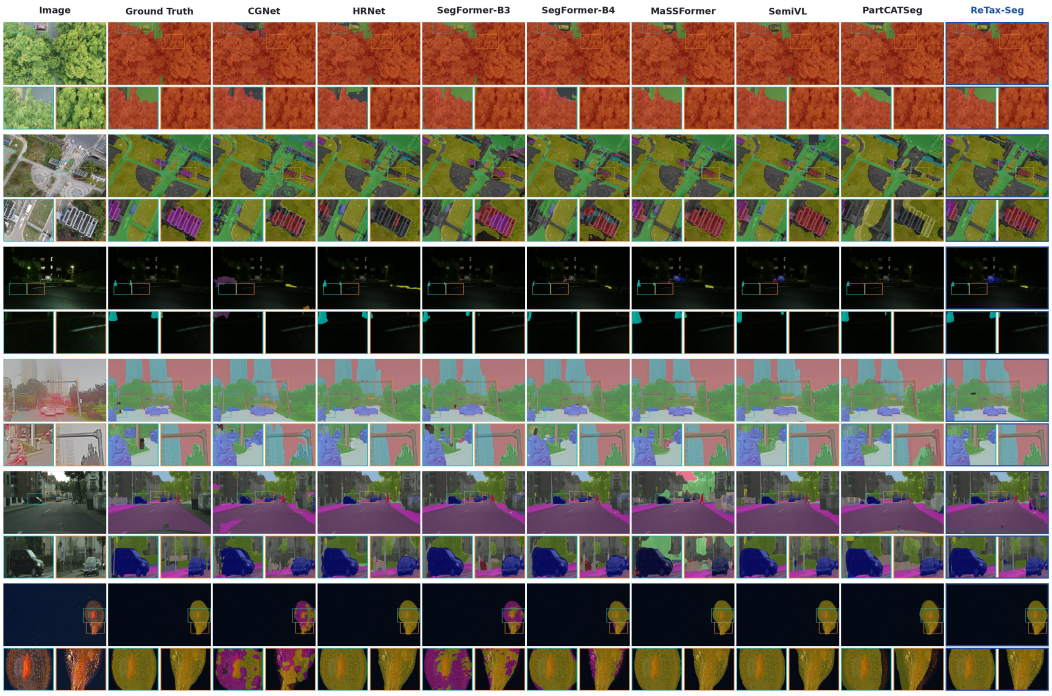

Figure 16: **Eight-method segmentation comparison with two enlarged details per image (set B).** The layout, method order, palettes, panel sizes, and enlargement protocol are identical to Figure 14. This second set uses disjoint images from the same six evaluation splits.

I Method-Specific Internal Diagnostics

We focus on four internal quantities that expose the actual computation of ReTax-Seg rather than generic backbone-energy or PCA views. Let x denote an image, and let r denote a decoder location. The value $p_\theta(z \mid x, r)$ is the predicted probability of atom z. The *atomic evidence* map is

$$E_{\text{atom}}(x, r) = \max_{z \in \mathcal{Z}} p_\theta(z \mid x, r), \tag{34}$$

where $\max$ selects the largest probability over the atom set $\mathcal{Z}$. The scalar $E_{\text{atom}}(x, r) \in [0, 1]$ measures how decisively the visual stream supports one atom. The *SAM reliability* map is the learned gate $\rho(x, r) \in [0, 1]$. It controls how strongly the frozen MobileSAM region feature contributes at each location. It is not a post-hoc saliency score.

For dataset d, native class c, atom z, and low-rank factor u, the region-conditioned correction is

$$\Delta_{d,r,c,z} = \sum_{u=1}^{S} q_{r,u} U_{d,c,u} V_{z,u}, \qquad S = 8. \tag{35}$$

The *dynamic transport* panel visualizes the root-mean-square magnitude of the spatially centered $\Delta_{d,r,c,z}$ over all native classes and atoms. The vector q_r is the region state. The matrices U_d and V are the dataset-class and atom factors. The panel highlights where the class–atom relationship changes relative to its image-wide value, rather than merely displaying foreground activation. Finally, let π_r be the forward atom posterior and let $\widehat{\pi}_r$ be its reverse reconstruction. The *cycle residual* is

$$R_{\text{cyc}}(x, r) = \tfrac{1}{2} \|\pi_r - \widehat{\pi}_r\|_1. \tag{36}$$

The subscript 1 denotes the ℓ_1 norm. Thus, $R_{\text{cyc}}(x, r)$ is half the absolute componentwise difference. It localizes disagreement between the forward atom-to-class mapping and its reverse reconstruction.

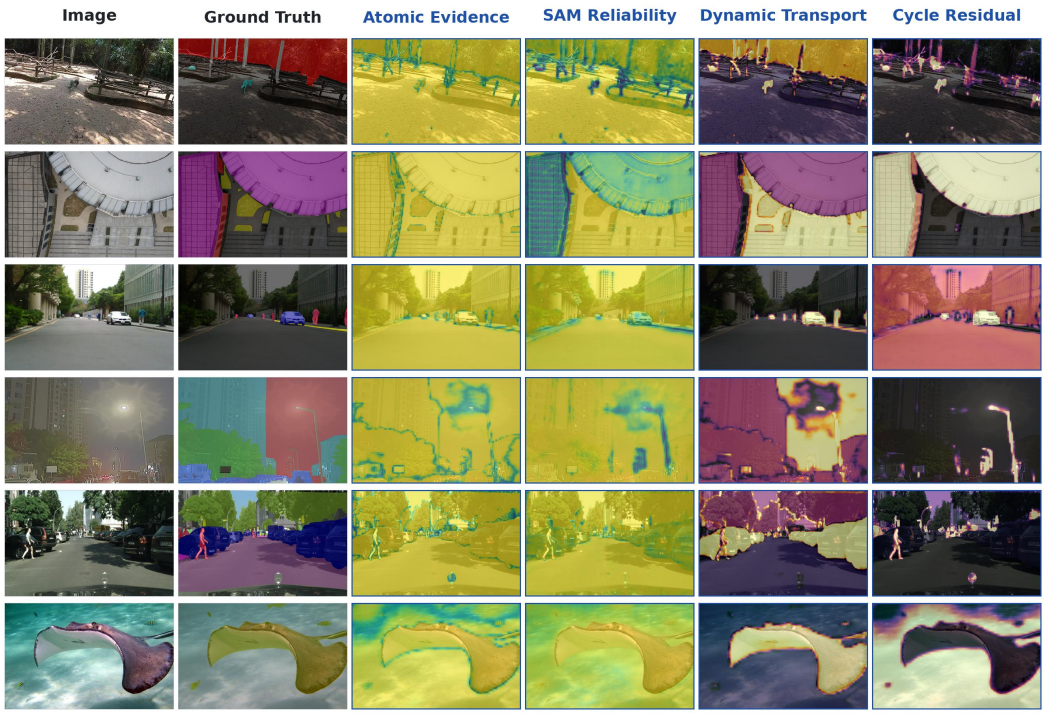

Figure 17: **ReTax-specific internal diagnostics on all six datasets.** Columns show the input, ground truth, atom evidence, learned SAM reliability, spatial variation of dynamic transport, and bidirectional cycle residual. Rows follow AeroScapes, VDD, MFNet, FMB, Cityscapes, and SUIM. Evidence and reliability use the fixed range $[0, 1]$. Transport and residual maps use a per-panel 2nd–98th percentile display range. The latter normalization is only for visualization and does not alter model inference.

The four views have complementary interpretations. Atomic evidence reveals where the shared semantic vocabulary is decisive. The reliability gate changes substantially across aerial, low-light, urban, and underwater data, which is consistent with selectively attenuating a frozen region prior when it is unreliable. Dynamic transport varies most visibly around boundaries, small objects, and ambiguous structures, while the cycle residual exposes locations at which the two mapping directions disagree. These maps directly audit the proposed modules, but they remain diagnostic rather than causal evidence. The component ablations in the main paper establish their quantitative contribution.

The gate statistics also vary across datasets rather than collapsing to one constant. On the displayed samples, median SAM reliability is 0.973 for AeroScapes, 0.772 for VDD, 0.973 for MFNet, 0.982 for FMB, 0.980 for Cityscapes, and 0.938 for SUIM. The lower VDD value is consistent with the model attenuating a frozen region prior under a substantial viewpoint and scale shift. High median reliability in another domain does not imply that every pixel trusts SAM: the spatial maps contain localized low-reliability areas, and only those local values multiply the SAM contribution.

Cycle residual behaves differently from the gate. Its median is below 0.006 for AeroScapes, FMB, and Cityscapes, but is 0.431 for SUIM, 0.497 for MFNet, and 0.848 for VDD on these samples. This range is useful evidence against a trivial identical operating point across datasets. It is not, however, a calibrated error probability: the residual measures disagreement between two latent mapping directions, and its absolute scale depends on the independently learned native taxonomy. We therefore interpret spatial concentration within a dataset and avoid ranking datasets by residual magnitude.

J PER-DATASET EFFECTIVENESS PROFILES

The aggregate plots are expanded so that every dataset has its own bar chart and radar profile. Figure 18 compares mIoU for exactly the same eight methods used in the qualitative figures. It preserves large differences in absolute difficulty across domains that would disappear in a single dataset mean. ReTax-Seg obtains 80.34, 84.50, 71.91, 60.64, 72.67, and 71.74% mIoU on AeroScapes, VDD, MFNet, FMB, Cityscapes, and SUIM, respectively.

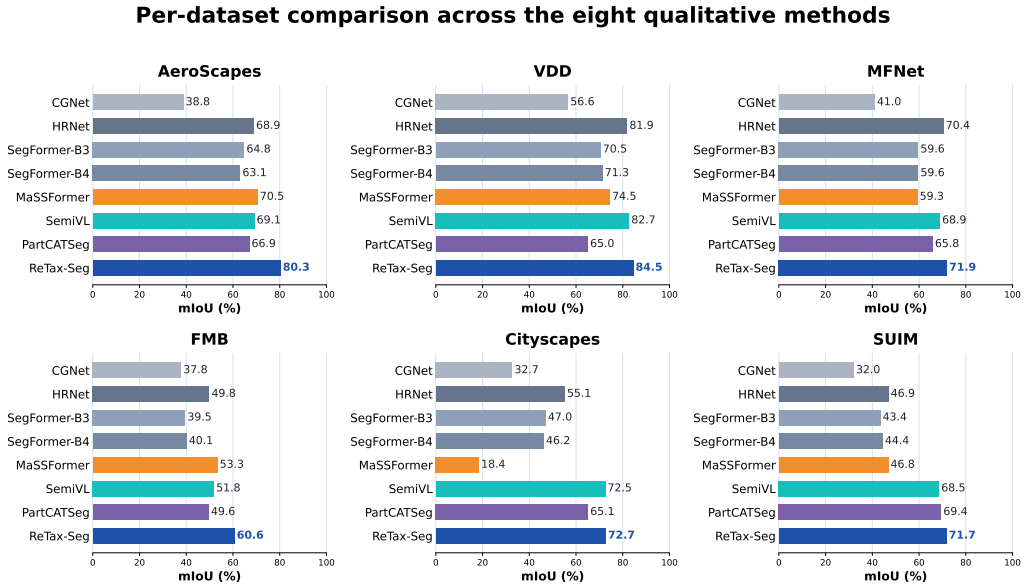

Figure 18: **One complete-split mIoU comparison per dataset.** Each panel reports CGNet, HRNet, SegFormer-B3, SegFormer-B4, MaSSFormer, SemiVL, PartCATSeg, and ReTax-Seg in a fixed order. Values come from complete saved predictions, not from the qualitative samples.

We select the strongest of all 20 alternatives independently on each dataset. Relative to those values, the mIoU margins are +1.28, +1.83, +1.22, +0.59, +0.17, and +2.32 percentage points in dataset order. Their mean is approximately +1.24 points, calculated from the displayed two-decimal table values. These comparisons use the full baseline set rather than only the seven alternatives shown in Figure 18. The largest advantage is on SUIM, followed by VDD. The smaller FMB and Cityscapes margins show that the gains remain positive when strong competing methods already perform well on the native taxonomy.

Absolute performance and relative improvement answer different questions. VDD has the highest ReTax-Seg mIoU (84.50%), whereas SUIM has the largest margin over its strongest alternative (+2.32 points) at an absolute score of 71.74%. FMB has the lowest absolute score (60.64%) and improves over SAM2-UNet by 0.59 points. This separation supports reporting both the score and the competitive margin for every dataset rather than a single average.

Figure 19 similarly provides one six-metric radar chart per dataset. We show ResNeSt-200-DLV2, SemiVL, PartCATSeg, and ReTax-Seg, the four representative methods for which complete values for all six metrics are available. No cross-dataset normalization is applied: every axis remains in percentage points on $[0, 100]$. The view shows that the improvement is not restricted to mIoU, although the six axes are correlated functions of the same confusion matrix and must not be interpreted as independent tests.

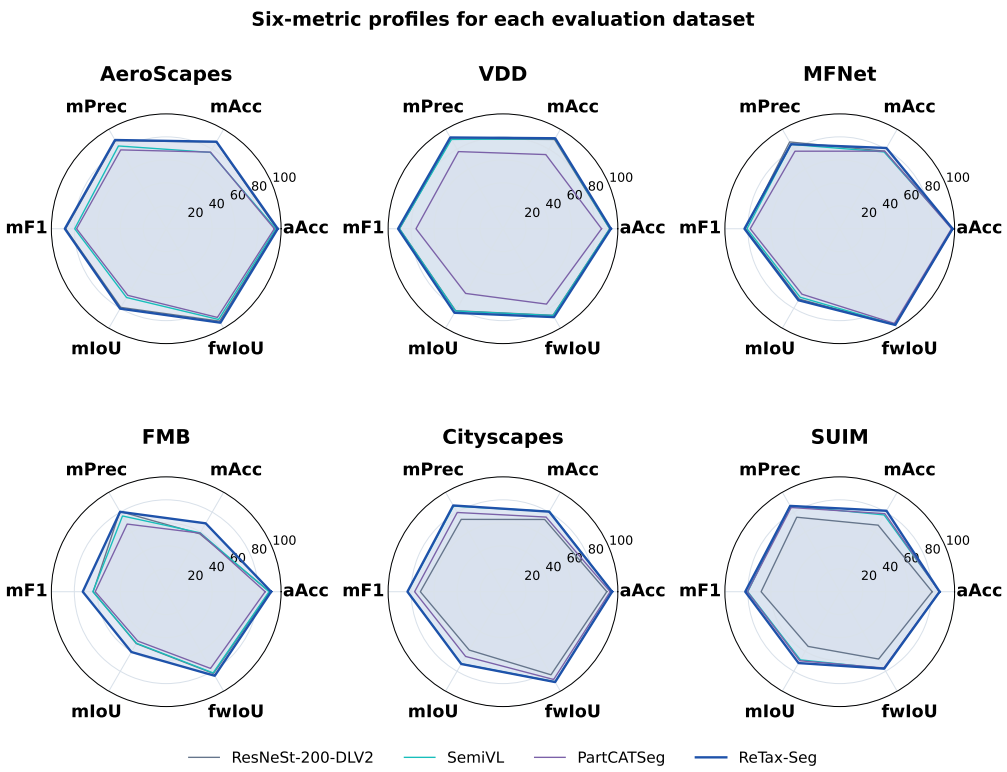

Figure 19: **One six-metric profile per dataset.** Axes are overall accuracy (aAcc), mean accuracy (mAcc), macro precision (mPrec), macro F1, mIoU, and frequency-weighted IoU (fwIoU). All panels use the same 0–100 scale and method colors.

The radar profiles rule out the narrow explanation that the gain comes only from one IoU aggregation. Against the strongest alternative in each cell, ReTax-Seg leads in 30 of 36 dataset–metric comparisons. The six exceptions are MFNet macro precision, FMB macro precision, Cityscapes mean accuracy, and SUIM overall accuracy, macro precision, and frequency-weighted IoU. In particular, the method wins mIoU on all six datasets while not uniformly maximizing precision. This is consistent with improving class recovery and overlap without obtaining every possible operating-point trade-off.

K Per-Dataset Confusion Structure

Figure 20 reports one complete-split confusion matrix for every independently trained ReTax-Seg model. Each row is normalized by the number of valid ground-truth pixels in that native class, so diagonal entries are class recall and off-diagonal entries identify the destination of errors. To keep the 15-class FMB and 19-class Cityscapes panels legible at a large type size, we print their diagonal values. Every off-diagonal cell remains encoded by the common color scale. The smaller matrices additionally print every off-diagonal entry above 2%.

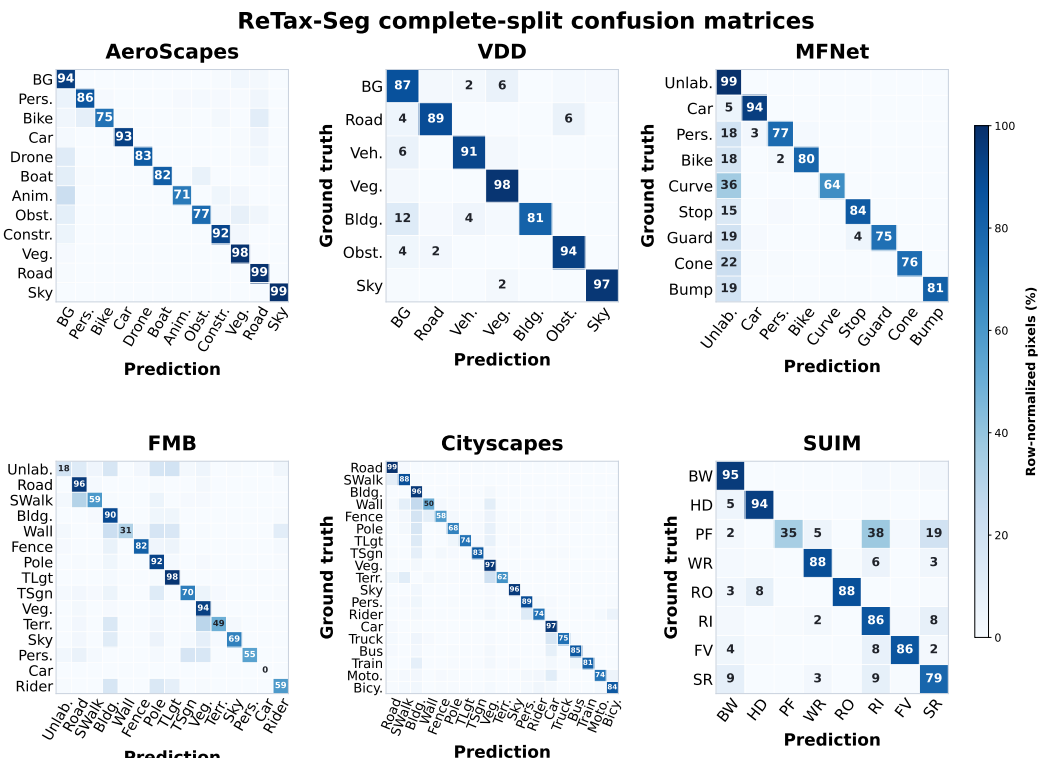

Figure 20: **Complete-split native-class confusion matrices for all six datasets.** Rows are ground truth and columns are predictions. Values and colors are row-normalized percentages. The panels use the best independent checkpoint and the same test split as the main quantitative table. Class names are abbreviated only to preserve a large, readable type size.

The matrices expose annotation-specific rather than universal error modes. In AeroScapes, animal has 71.3% recall and 22.0% of its pixels are assigned to background. VDD building has 80.7% recall with 12.2% assigned to background. MFNet concentrates several errors in its broad unlabeled class. The model sends 36.0% of curve pixels and 22.2% of color-cone pixels there. It also sends 19.5% of guardrail pixels there. These destinations are meaningful because "unlabeled" is an annotation policy. It is not a coherent visual concept that can be recovered from its name alone.

The denser taxonomies show a different structure. In FMB, sidewalk is most often confused with road (30.8%), terrain with vegetation (28.5%), and wall with building (21.9%). The displayed zero for car corresponds to zero valid car pixels in this evaluation split and is not a measured zero-recall failure. Cityscapes wall and fence reach 50.1% and 57.9% recall, with building as their largest error destination. SUIM's PF class is the principal outlier at 34.6% recall, with 37.5% assigned to RI and 19.2% to SR. These cases identify where finer visual evidence or a richer atom vocabulary may be more valuable than further sharpening the current transport.

L Directed Cross-Dataset Transfer

Figure 21 expands the averaged cross-dataset result in Table 7 into all 30 directed transfers for ReTax-Seg. A row fixes the training dataset and a column changes the unseen target taxonomy. No target image or annotation is used for optimization. Transfer is asymmetric, as expected when source domains and annotation policies differ. Cityscapes-trained features transfer most broadly in this experiment, while targets with severe sensing or policy shifts remain hard. Over all directed off-diagonal cells, ReTax-Seg reaches 9.45% mean mIoU, compared with 9.07% for the strongest reported comparator, HQ-SAM. Several domain pairs nevertheless remain close to zero, so the gain should not be read as solved universal zero-shot segmentation.

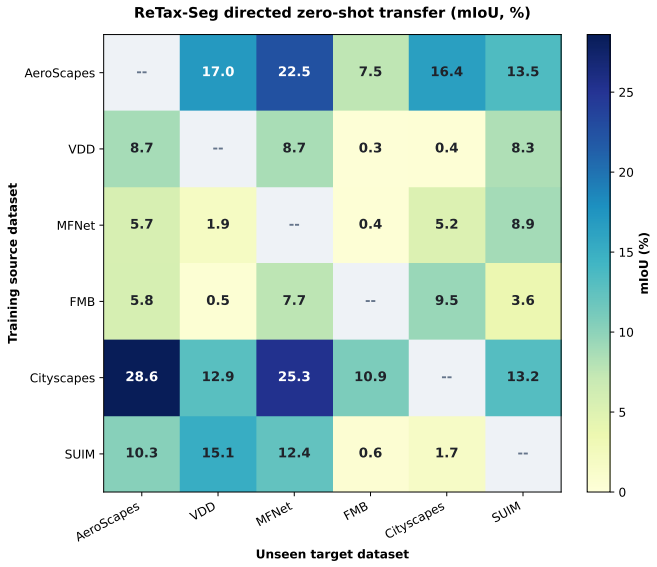

Figure 21: **Directed zero-shot transfer matrix of ReTax-Seg.** Rows denote the training source and columns the unseen target. Entries are complete-split mIoU in percent with no target fine-tuning. The diagonal is omitted because it is in-domain evaluation.

Direction matters substantially. Training on Cityscapes and testing on AeroScapes gives 28.6% mIoU, whereas the reverse direction gives 16.4%. The same asymmetry appears for several other pairs and cannot be explained by target difficulty alone. A source taxonomy determines which distinctions the backbone and compiler learn to preserve. A target can only reuse distinctions that remain represented. Thus, cross-dataset transfer tests the paper's shared semantic organization claim more directly than six independent in-domain results, while also showing that taxonomy mismatch is not solved.

Cross-taxonomy protocol. For a conventional native head, the source logits cannot be evaluated in a target label space without an explicit conversion. We construct that conversion once, before evaluation, by maximizing definition overlap through the same 39-atom vocabulary and then keep it fixed for every target image. ReTax-Seg instead requests the target compiler stored in the source checkpoint. Target-specific parameters receive no target-image update and remain at their initialized semantic prior. The learned visual atom representation and shared atom factors come from the source run. The comparison therefore tests a fixed semantic map against the proposed factorized compiler. It is not target fine-tuning or prompt selection.

Every heat-map cell aggregates all valid pixels in the canonical target split. The diagonal is withheld because its training and evaluation taxonomies match and it would be orders of magnitude easier than the off-diagonal problem. Near-zero cells are retained and included in the mean. This matters for FMB as a target, where several source models recover little usable overlap, and prevents successful related-domain transfers from hiding incompatible pairs.

M Further Quantitative Diagnostics

The following four views summarize complementary properties of the measured results. They are derived from the same complete-split metric files, directed transfer runs, raw confusion matrices, and saved 39-dimensional atom probabilities used elsewhere in this supplement. Thus, they add no selected images or separately tuned evaluation protocol.

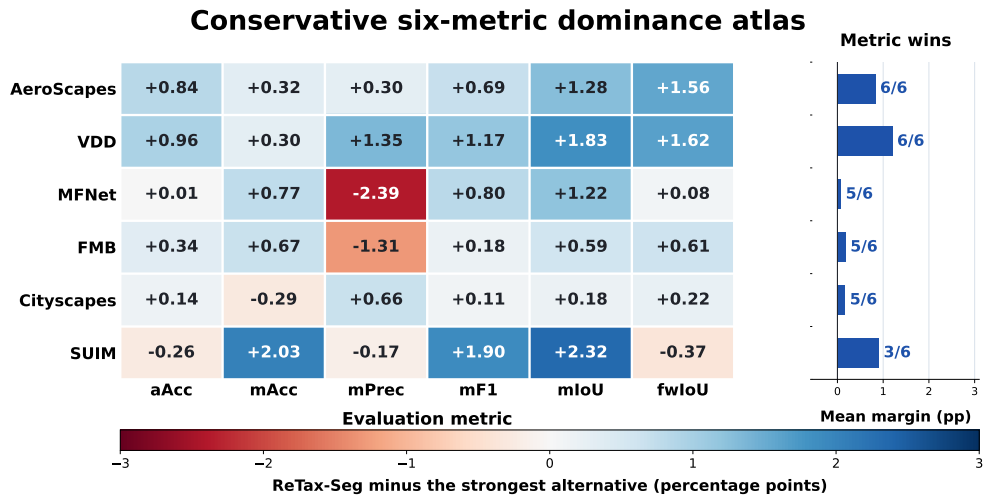

Figure 22: **Conservative metric-dominance atlas.** Each heat-map entry is the score of ReTax-Seg minus the strongest of all 20 alternatives for that specific dataset–metric pair. This is stricter than comparison with one fixed baseline. Blue cells favor ReTax-Seg and orange cells favor the best alternative. The marginal bars report the mean signed margin and number of wins. ReTax-Seg leads in 30 of the 36 comparisons, including mIoU on all six datasets.

The mean signed margins over the strongest alternatives are +0.83 points for AeroScapes, +1.20 for VDD, and +0.08 for MFNet. They are +0.18 for FMB, +0.17 for Cityscapes, and +0.91 for SUIM. Averaged across all 36 cells, the conservative margin is +0.56 points. The negative cells are also retained rather than hidden: the largest is MFNet macro precision at -2.39 points, which indicates that improved overlap need not maximize the per-class precision operating point.

The six columns emphasize different behavior. Overall accuracy and frequency-weighted IoU are dominated by common pixels. Mean accuracy, macro precision, macro F1, and mIoU give rare native classes much more influence. Leading on both groups is stronger evidence than improving only a frequency-weighted score. At the same time, these quantities all derive from one confusion matrix and are statistically dependent. We therefore use the atlas as a behavior profile, not as 36 independent significance tests.

The exceptions are diagnostically consistent. MFNet and FMB lose macro precision while improving macro F1 and mIoU, indicating a recall–precision trade-off rather than a collapse of common classes. Cityscapes loses 0.29 points in mean accuracy but gains on the other five metrics. SUIM loses small amounts of overall accuracy, macro precision, and frequency-weighted IoU while improving mean accuracy, macro F1, and mIoU. These signed patterns are more informative than reporting only that a row is “best” on average.

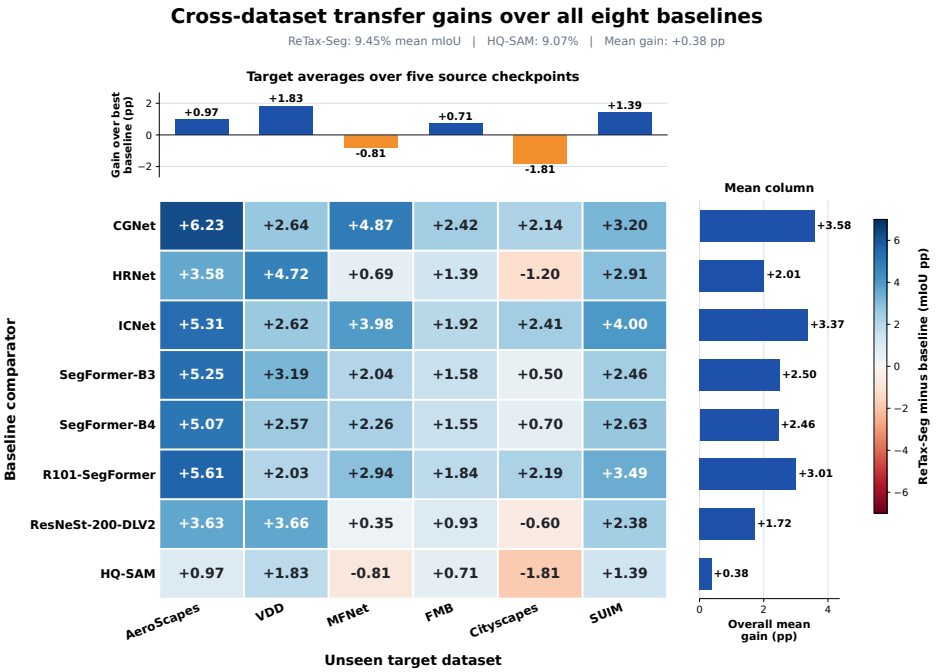

Figure 23: **Cross-dataset transfer gains against all eight baselines.** Each cell subtracts a baseline's target-averaged mIoU in Table 7 from the corresponding ReTax-Seg value. Each target score averages five foreign-source checkpoints. Top bars compare against the strongest baseline for each target. Right bars subtract the reported overall mean columns. All differences use the displayed two-decimal table values, keeping the comparison aligned with the main paper.

Figure 23 uses exactly the eight baselines in Table 7, including HQ-SAM. ReTax-Seg has the highest overall mean mIoU at 9.45%, compared with 9.07% for HQ-SAM, a gain of 0.38 percentage points. It exceeds every baseline in the overall mean, while the target-wise comparison exposes exceptions that this average conceals.

Against the strongest baseline for each target, the gains are +0.97, +1.83, -0.81, +0.71, -1.81, and +1.39 points for AeroScapes, VDD, MFNet, FMB, Cityscapes, and SUIM, respectively. HQ-SAM is the strongest alternative for all six target averages. Thus, ReTax-Seg leads on four targets, with its largest gain on VDD, but trails HQ-SAM on MFNet and Cityscapes. The latter deficits are retained to make the limits of cross-taxonomy transfer explicit.

The heat map distinguishes comparison against a fixed architecture from comparison against the strongest alternative on each target. Each cell summarizes five source checkpoints, so it does not establish how many of the 30 individual directed transfers outperform a competitor. Source-specific asymmetry remains visible in Figure 21. The present figure provides a complementary comparison at the same aggregation level as the main cross-dataset table.

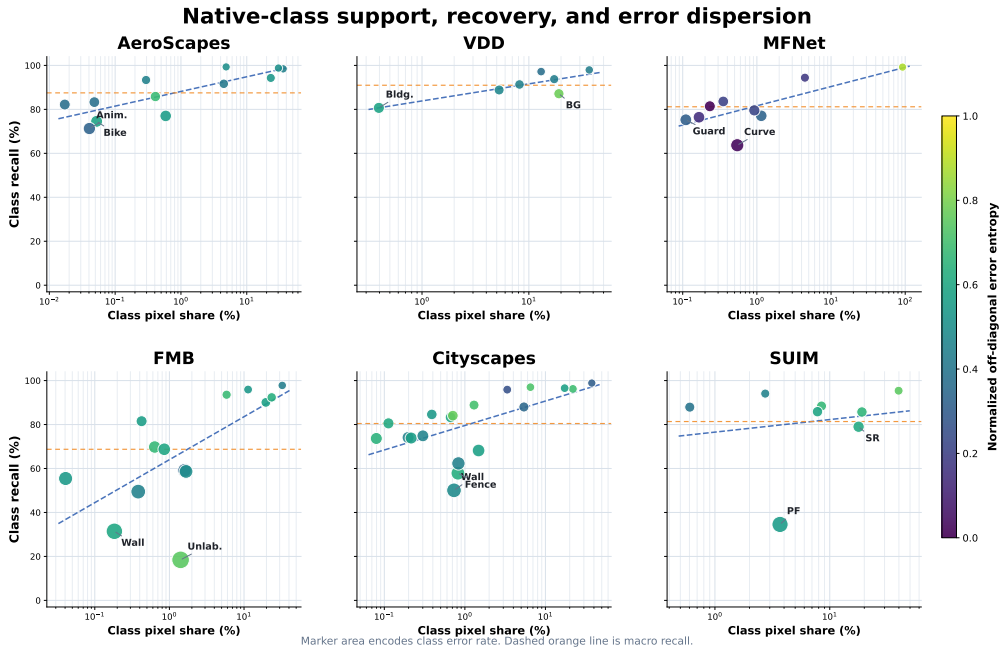

Figure 24: **Native-class support–error landscape.** Every point is one native class from the complete evaluation split. Horizontal position is its pixel share, vertical position is recall, marker area is error rate, and color is normalized entropy of the off-diagonal error destinations. The dashed blue fit summarizes the support–recall trend, while the orange line is the dataset macro recall. Labels identify the two lowest-recall classes in each dataset. The plot distinguishes rare-class failure from diffuse taxonomy confusion without collapsing either into one scalar metric.

Support explains part, but not all, of the class-level variation. The fitted trend is positive in each panel, yet classes with similar pixel share can have substantially different recall and error entropy. Low entropy means that most mistakes flow to one competing label and suggests a specific taxonomy ambiguity. High entropy indicates diffuse failure across several destinations. This distinction matters for method development: a concentrated wall-to- building error calls for better class–atom discrimination, whereas diffuse small-object errors may require stronger spatial evidence or sampling.

For class c, the plotted entropy is computed after removing the diagonal and renormalizing the remaining row, then dividing Shannon entropy by $\log(C_d - 1)$. It therefore lies in $[0, 1]$ independently of the number of native classes. Here recall_c is the recall of class c. Marker area encodes $1 - \mathrm{recall}_c$, not class support. Support is already shown on the logarithmic horizontal axis. Classes with no valid evaluation pixels are omitted from the fitted trend and should not be interpreted from a displayed zero diagonal.

The landscape also separates two versions of the long-tail problem. A rare class at high recall shows that frequency alone is not limiting. A frequent class at low recall points instead to label ambiguity or insufficient discrimination. FMB wall and Cityscapes wall/fence are examples of the latter, while several tiny AeroScapes categories remain above 75% recall. This is one reason the paper focuses on dynamic taxonomy interpretation in addition to generic small-object modeling.

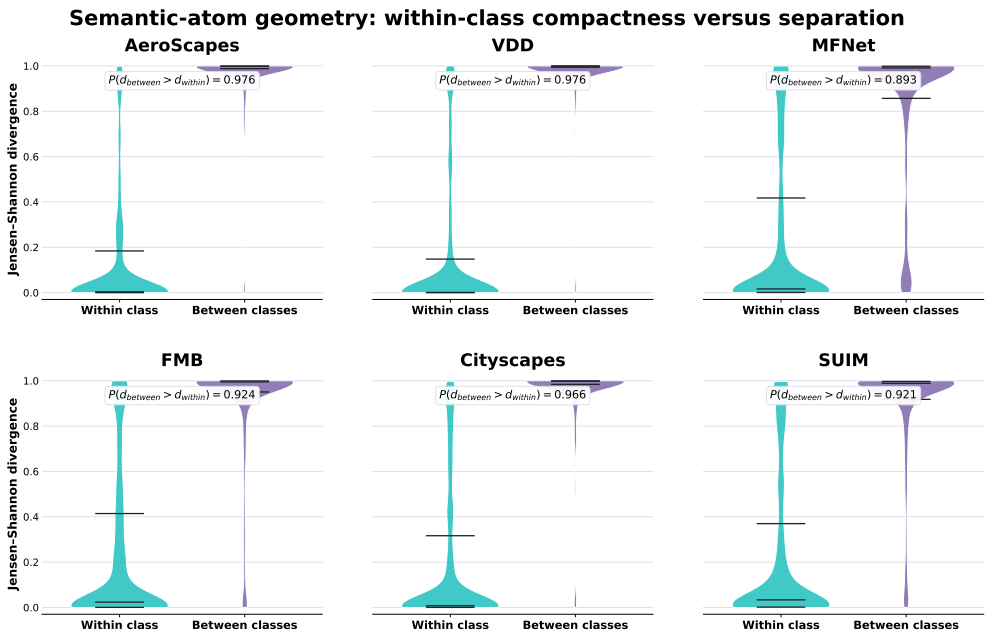

Figure 25: **Semantic-atom geometry audit before two-dimensional projection.** For every dataset, the violins compare Jensen–Shannon divergence between pairs of real 39-dimensional atom distributions from the same native class and from different native classes. Horizontal lines mark quartiles. The inset probability is the chance that a randomly sampled between-class distance exceeds a within-class distance. Values from 0.893 to 0.976 show substantial class structure in the atom space. Unlike t-SNE, this analysis is computed in the original representation and does not rely on a projected layout.

The separation probability is 0.976 on AeroScapes and VDD, 0.966 on Cityscapes, 0.924 on FMB, 0.921 on SUIM, and 0.893 on MFNet. All values are well above the 0.5 chance level, but none is treated as proof of a perfectly disentangled ontology. The long tails of both violins are important: some same-class pixels use different atom mixtures, and some different-class pixels remain close. Dynamic transport is intended precisely for this residual ambiguity. Forcing atom features into one rigid class prototype would discard the observed intra-class structure.

Jensen–Shannon divergence is used because every atom vector is a probability distribution and the measure is symmetric, finite when components approach zero, and bounded by one with base-2 logarithms. The reported probability is an ordering statistic: it asks how often a randomly drawn between-class pair is farther apart than a randomly drawn within-class pair. It is invariant to the t-SNE layout and avoids choosing a linear classifier, but it still depends on the deterministic image and pixel sample described in the protocol.

Ground-truth classes enter only in forming the two pair sets. The atom vectors themselves are produced before the native transport is applied, so separation cannot be attributed to reading the final class logits. Conversely, strong separation does not imply that atom indices have perfectly human-readable meanings. It shows that the shared probability space retains information that correlates with native classes, which is the weaker and relevant requirement for reversible compilation.

N Additional Internal Feature Visualizations

We next visualize the learned representation from complementary projections. All figures use cached tensors from real forward passes of the corresponding independent checkpoint. Ground truth is used only to group features or draw reference boundaries after inference. It is never supplied to the model.

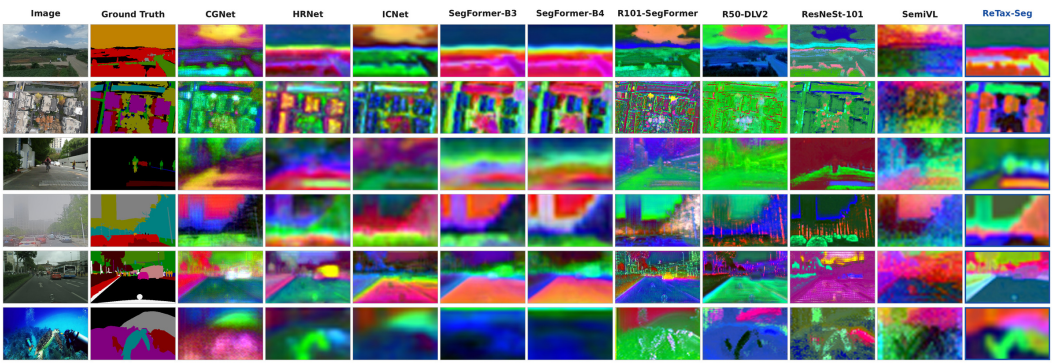

Figure 26: **PCA visualization of learned feature structure across all six datasets.** The 6×12 grid applies the same extraction and display pipeline to the independently trained best-validation checkpoints. RGB encodes the leading principal directions of the cached feature tensor. Colors are relative coordinates in the projected feature space, not semantic labels or comparable physical units across checkpoints.

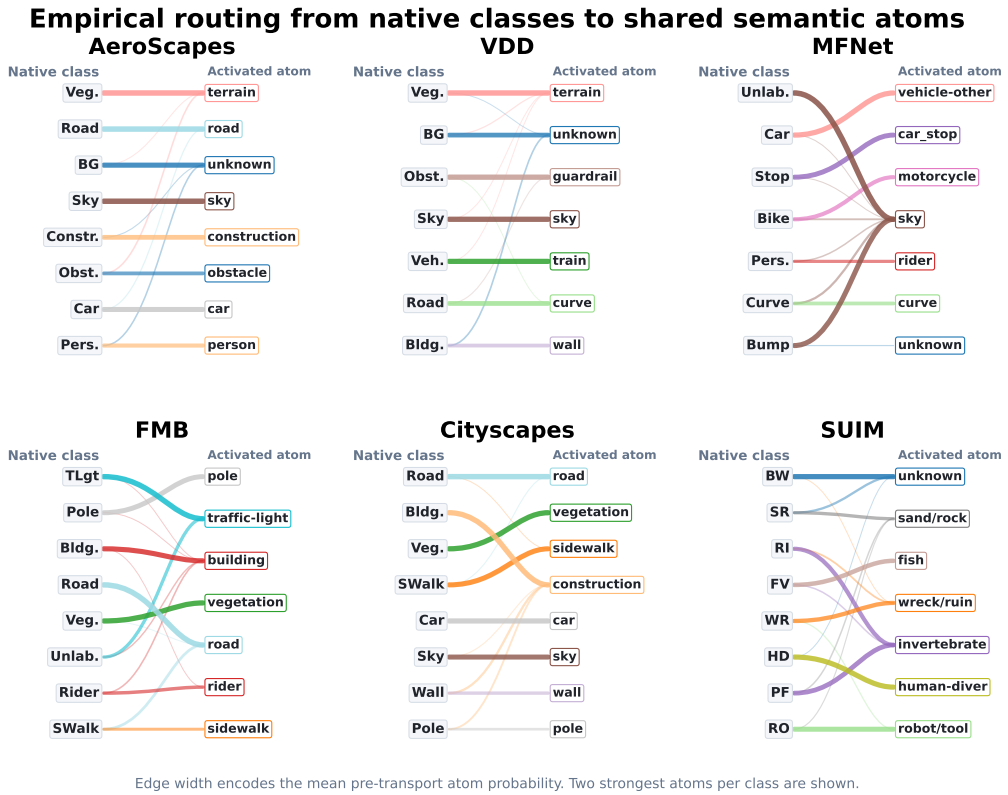

Figure 27: **Empirical native-class to semantic-atom feature routing.** For each of the eight most frequent sampled native classes (seven where fewer are present), we average the real 39-dimensional pre-transport atom probability and draw its two strongest components. Edge width is proportional to the mean probability. The view exposes both stable relations, such as road-to-road, and deliberately nonliteral visual decompositions that the region-conditioned compiler must map into the native taxonomy. It therefore shows learned feature use rather than the manually initialized compatibility mask.

Several routes are literal—road to road in AeroScapes and Cityscapes, sky to sky, and the SUIM underwater categories to domain-specific atoms. Others are compositional. Cityscapes building activates construction, VDD vehicle activates a vehicle-related train atom, and MFNet person distributes mass over rider and scene-dependent alternatives. These are pre-transport visual responses, not final class predictions. Their nonliteral cases illustrate why the text prior must remain soft: a hard name match would forbid useful visual atoms, while the learned compiler can recombine them into the supervised native class.

The diagram should also not be read as a causal graph. It summarizes the two largest components of a 39-dimensional distribution over a deterministic sample, so thin unshown routes still carry probability mass. We provide the full vectors in the released archive, and the separation analysis uses all 39 dimensions. The sparse drawing is used only to make the dominant routing pattern legible at paper scale.

The empirical route and the learned transport are complementary objects. The former answers which visual atoms activate inside pixels belonging to a native class. The latter answers how the compiler converts any activated atom into native-class probability at a particular location. A native class can therefore receive evidence from an atom that is not its literal name, and the same atom can contribute differently across regions. Conflating these two directions would reduce the model to the static lookup that the proposed compiler is designed to avoid.

The routing diversity is also dataset dependent. AeroScapes and Cityscapes contain several direct surface and structure routes, FMB mixes road-scene and actor evidence, and SUIM uses the domain-specific tail of the 39-atom vocabulary. MFNet shows the largest use of alternatives such as sky or vehicle-other for broad labels. These differences are compatible with one shared atom inventory serving multiple annotation policies without requiring identical per-class prototypes.

The fixed RGB semantics make cross-row visual comparison possible without claiming that the raw magnitudes are calibrated across checkpoints. Smooth areas are dominated by stable evidence and reliability, whereas narrow color changes occur near small objects and boundaries. The discontinuity column is derived jointly from all four signals, so it reveals locations at which the method changes its internal decision process even when no single heat map is dominant. The aerial VDD sample emphasizes large structural transitions. MFNet and FMB emphasize vehicles, people, and illumination boundaries. SUIM preserves the articulated animal contour under a very different scene appearance.

The final column provides a visual audit rather than a boundary metric. Cyan features that do not overlap orange ground-truth boundaries can correspond to texture or to a latent semantic change inside one annotated region. Conversely, an orange boundary without a cyan response indicates a location where the combined internal representation changes weakly. White overlap is therefore useful evidence of alignment, but the complete segmentation metrics remain the appropriate measure of output quality.

AeroScapes, MFNet, FMB, and Cityscapes show the clearest decay of atom uncertainty away from a native boundary. SAM attenuation often follows the same direction but decreases more slowly, indicating that the gate responds to region reliability rather than merely copying an edge detector. Transport variation is comparatively persistent in AeroScapes and rises away from the boundary in FMB and SUIM, while cycle residual remains high over broad regions in VDD, MFNet, and SUIM. These differences agree with the spatial composites: the compiler performs both boundary-local adjustment and region-level taxonomy adaptation.

Because each curve uses one fixed visualization sample and per-sample robust normalization, the profile is descriptive rather than a population estimate. The interquartile band shows spatial heterogeneity, not uncertainty over training runs. A future multi-seed study could aggregate these profiles over the complete split. Here they serve the narrower purpose of verifying that the saved latent maps are spatially structured and that their four mechanisms are not redundant copies of one response.

Combined interpretation. Taken together, Figures 17, 25, 27 and 28 test three distinct failure modes of a semantic compiler. The atom-space views test collapse, the routing view tests whether native classes use more than literal name matches, and the spatial views test whether transport and SAM influence vary with the image. No single visualization establishes causality, but the agreement of these independent diagnostics with complete-split gains makes a trivial static-lookup explanation less plausible.

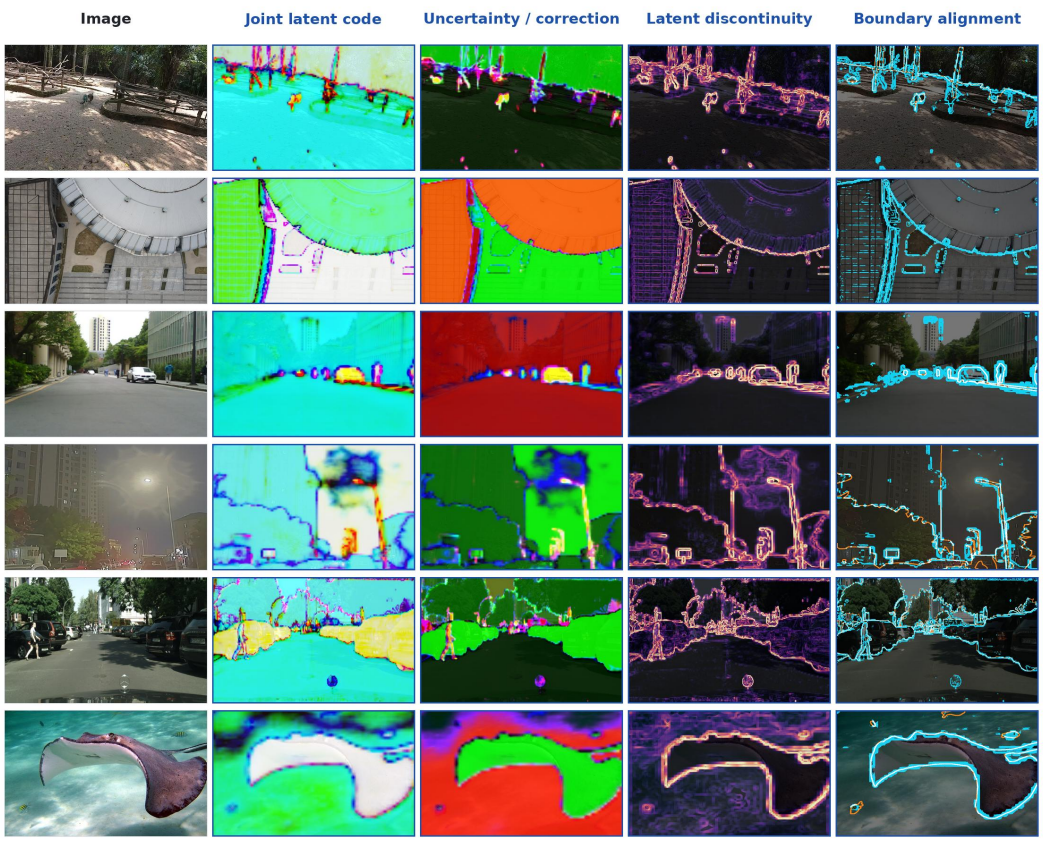

Figure 28: **Joint spatial encoding of method-specific latent features.** Rows follow AeroScapes, VDD, MFNet, FMB, Cityscapes, and SUIM. In the joint latent code, RGB denotes dynamic-transport variation, atom evidence, and SAM reliability. In the uncertainty/correction view, RGB denotes cycle residual, transport variation, and atom uncertainty. Latent discontinuity is the RMS spatial gradient over all four normalized internal maps. In the last column, the top decile of this discontinuity is cyan, native-label boundaries are orange, and their overlap is white. The same channel semantics are used in every row.

Limits of feature evidence. All spatial maps in this section come from one deterministic sample per dataset, selected before plotting from the same fixed positions as qualitative set A. They demonstrate that the implemented tensors have the stated spatial behavior. They do not estimate how frequently a pattern occurs over the full split. Likewise, t-SNE and routing use eight deterministic images and a capped number of pixels per class. The complete-split metrics and confusion matrices remain the basis for performance claims.

Feature color is not a universal physical unit. Evidence and reliability are probabilities, but transport variation and cycle residual have dataset- and checkpoint-dependent distributions. Whenever multiple signals are placed in one RGB image or profile, we disclose and fix the normalization. Colors can be compared spatially within a panel and semantically across panels.