VIRA: SEMANTIC ATTRIBUTE-GUIDED ADAPTATION FOR CONTINUAL OPEN-VOCABULARY SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary segmentation (OVS) must incorporate newly observed data while preserving its ability to respond to changing query vocabularies. However, continual fine-tuning can couple adaptation to the training vocabulary and degrade performance across other domains and class sets. We introduce VIRA (*Vocabulary-Induced Relation-Guided Adaptation*), which represents runtime classes through shared semantic attribute coordinates. Class–attribute relations are grounded with region-level visual evidence and converted into a fixed-dimensional condition that controls low-rank residual composition. For continual deployment, completed adaptations are frozen and retained, while compatibility retrieval and region-wise arbitration determine when to use them or fall back to the frozen base. We evaluate VIRA on a COCO-pretrained $\rightarrow$ A-150 $\rightarrow$ Cityscapes $\rightarrow$ Mapillary Vistas stream, together with retention benchmarks, expanded taxonomies, and cross-architecture transfer. Capacity-matched controls, coordinate-inventory comparisons, mechanistic interventions, and routing diagnostics examine the role of semantic attributes and deployment cues. Across the three incremental domains, VIRA improves macro mIoU/PQ by +4.33/+2.84 over the frozen source, while preserving access to frozen-source behavior on retention benchmarks.

1 INTRODUCTION

Open-vocabulary segmentation (OVS) extends segmentation beyond a fixed label space through text-derived category representations, allowing the queried vocabulary to vary at inference time (Radford et al., 2021; Xu et al., 2023; Cho et al., 2024; Xie et al., 2024; Li et al., 2025). This flexibility is important in deployment, where both visual distributions and queried concepts may change over time. In such settings, models may need to incorporate observed data through continual fine-tuning while preserving the open-vocabulary recognition capability of the pretrained model. Maintaining this capability during continual adaptation is therefore a challenge for continual OVS.

However, fine-tuning on newly observed data can conflict with preserving the model's open-vocabulary capability. Parameter updates learned from one dataset may alter representations that previously supported recognition across other domains and query vocabularies, leading to degradation outside the current training distribution. This issue is important in OVS, where inference may involve enlarged, merged, or otherwise changed query sets. If adaptation becomes overly coupled to the vocabulary associated with the training data, the resulting model may lose robustness to other domains. The central challenge is therefore not merely to support variable-size vocabularies, but to learn new adaptation while protecting the model's ability to respond to open-vocabulary queries.

Addressing this challenge requires a control representation that can exploit query semantics without being tied to a class set. Region-only conditioning does not use the queried vocabulary and therefore cannot exploit query semantics; globally pooling query embeddings supports variable-size inputs but weakens class-specific structure; and class-specific adapters retain class-specific control but make adaptation capacity depend on queried classes. We therefore seek a representation reusable across vocabularies, preserves class-specific semantics, and admits region-level visual grounding. Semantic attributes offer a natural candidate: different classes can share properties such as material, geometry, appearance, and state, while their class–attribute relations remain class-specific. Because many attributes also admit region-level visual evidence, the same coordinates can be grounded in local

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.

image content. Semantic attributes thus provide a shared coordinate system that can remain reusable across changing vocabularies while retaining class structure and local visual relevance.

Motivated by this observation, we introduce VIRA (*Vocabulary-Induced Relation-Guided Adaptation*). VIRA maps class embeddings to sparse class–attribute relations over a shared frozen attribute inventory, then combines frozen-base class probabilities with region-level visual evidence in the coordinate space. The fixed-dimensional condition controls low-rank residual composition for region-specific adaptation. Within a specialist, vocabulary changes affect class-dependent computation without changing downstream dimensionality. For continual deployment, adaptations are frozen; compatibility retrieval selects a module, and region-wise arbitration selects between it and the frozen base. This separates attribute-guided adaptation from module selection and deployment.

We evaluate VIRA on a COCO-pretrained $\rightarrow$ A-150 $\rightarrow$ Cityscapes $\rightarrow$ Mapillary Vistas stream, with retention benchmarks, expanded taxonomies, and transfer to another OVS architecture. Experiments include capacity-matched conditioning controls, attribute and non-attribute coordinate inventories, mechanistic interventions, vocabulary perturbations, routing/arbitration diagnostics, and cross-architecture transfer. These studies test whether gains arise from semantic attribute structure rather than generic query access or target relevance, how shared coordinates participate in residual adaptation, and whether deployment depends on dataset-specific vocabulary cues.

Our contributions are threefold:

- A semantic attribute-coordinate representation that provides a reusable space across changing vocabularies while preserving class-specific semantics and supporting visual grounding.
- VIRA, which combines class–attribute relations, visual evidence, and low-rank residual composition to produce query-aware, region-specific adaptation.
- A modular continual OVS framework with frozen adaptation modules, compatibility retrieval, and region-wise arbitration, with diagnostics that isolate the role of semantic attributes.

2 Related Work

Open-vocabulary segmentation and semantic coordinates. OVS transfers vision–language representations to dense prediction through region–text alignment, cost aggregation, efficient encoder–decoders, and mask-level adaptation (Radford et al., 2021; Xu et al., 2023; Cho et al., 2024; Xie et al., 2024; Li et al., 2025), with recent work exploring semantic guidance, prompt-conditioned costs, spatial CLIP features, and retrieval-supported recognition (Wang et al., 2025; Zhao et al., 2025; Sun et al., 2025; Xuan et al., 2025; Aravanis et al., 2026; Kombol et al., 2026). Attribute learning provides cross-category structure (Jayaraman & Grauman, 2014; Pham et al., 2021; Bravo et al., 2023; Chen et al., 2023; Ramanathan et al., 2023), as prediction targets or recognition cues. VIRA instead uses frozen attribute text embeddings as control coordinates, with matched non-attribute and random-coordinate controls separating attribute structure from target relevance and text-space effects.

Continual OVS and parameter-efficient adaptation. Continual segmentation uses distillation, replay, regularization, pseudo-labeling, and modularization to limit forgetting (Cermelli et al., 2020; Douillard et al., 2021; Cermelli et al., 2023; Gong et al., 2024), with universal/open-world and continual test-time extensions (Yin et al., 2025; Lin et al., 2025; Lai et al., 2026). Continual OVS includes frozen-backbone retrieval (Gui et al., 2024), input-distribution-based expert combination (Hwang et al., 2025), and growing support sets (Aravanis et al., 2026). LoRA provides low-rank parameter-efficient updates (Hu et al., 2022), while conditional and mixture-of-adapter methods modulate fixed residual banks from context. VIRA instead uses shared semantic coordinates for query-aware within-specialist control, retaining class-specific structure and supporting regional visual grounding, while separating specialist retrieval and base–specialist arbitration from adaptation.

3 Method

3.1 Continual OVS setting

We start from a pretrained OVS model f_{θ_0} and sequentially observe domains $\mathcal{D}_1, \ldots, \mathcal{D}_T$, with $\mathcal{D}_t = \{(x_n, y_n, \mathcal{V}_t)\}_{n=1}^{N_t}$. Historical images are not replayed. At inference, the model receives an

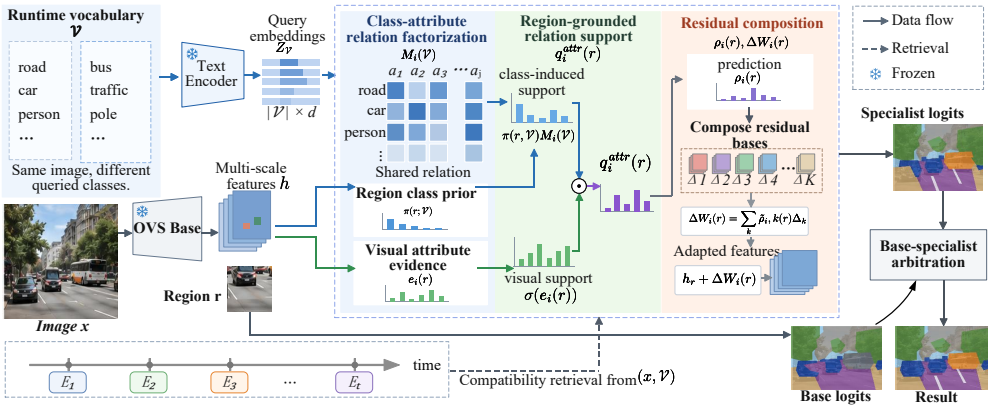

Figure 1: **Semantic attribute-coordinate interface.** Runtime queries and regional evidence are expressed in shared coordinates whose region-specific condition controls low-rank residual composition. The lower strip summarizes specialist acquisition, retrieval, and base–specialist arbitration.

image x and query vocabulary $\mathcal{V}$, but no explicit task identifier. After increment t, the system contains the frozen base, shared fixed components, a specialist bank $\mathcal{B}_t = \{E_i\}_{i=1}^{t}$, and frozen retrieval statistics $\mathcal{K}_t$. Completed specialists are never updated again.

Inference runs the base and one retrieved specialist, then arbitrates between their regional outputs. We separate *which specialist to use* from *how that specialist adapts*: retrieval/arbitration handle the former, while the vocabulary-to-adaptation interface below handles the latter.

3.2 Vocabulary-to-adaptation interface

For specialist E_i, VIRA performs three steps (Figure 1): map query classes into a shared semantic attribute basis, ground the class-induced coordinates with regional visual evidence, and use the resulting J-dimensional condition to compose low-rank residual bases. The query set participates in both the class-relation and regional class-evidence computations, while the downstream controller retains a fixed dimensionality.

Design rationale. The role of the attribute basis is not merely dimensional compatibility: generic pooling can already map a variable-size query set to a fixed-size vector, but it weakens explicit class structure, whereas class-specific controllers tie adaptation capacity to the queried classes. VIRA instead represents each class through relations to reusable semantic coordinates and grounds the same coordinates with local visual evidence. This yields a query-aware, class-structured regional condition that can be reused across changing vocabularies; fixed J-dimensional control is a consequence of this shared coordinate representation rather than the primary objective.

Representations. The frozen base provides region features $h_r \in \mathbb{R}^d$ and normalized query embeddings $Z_{\mathcal{V}} \in \mathbb{R}^{|\mathcal{V}|\times d}$. A shared matrix $A \in \mathbb{R}^{J\times d}$ stores J fixed normalized coordinates; by default, $J = 52$ attribute text embeddings. Relation and coefficient networks are specialist-specific, whereas A and the base representations are shared. Alternative inventories are studied in Section 4.4; implementation details are in Appendix A.

3.3 Class level: projecting queries into attribute coordinates

Specialist E_i learns low-rank factors $P_i, Q_i \in \mathbb{R}^{d\times r_a}$ and computes

$$M_i(\mathcal{V}) = \mathrm{Entmax}_{1.5}\left(\frac{Z_{\mathcal{V}}(I + P_i Q_i^\top)A^\top}{\tau_a}\right) \in \mathbb{R}^{|\mathcal{V}|\times J}. \tag{1}$$

Entmax acts independently on each class row (Peters et al., 2019); we use $r_a = 64$. The identity term preserves the frozen text-similarity geometry, while $P_i Q_i^\top$ learns a specialist-specific correction.

Adding a query class appends a row to $M_i(\mathcal{V})$ without adding parameters, retaining class-specific relation structure while keeping the learned relation correction independent of vocabulary size.

3.4 Region level: grounding coordinates in visual evidence

The frozen base induces regional class probabilities

$$\pi(r;\mathcal{V}) = \mathrm{softmax}(\kappa \bar{h}_r Z_{\mathcal{V}}^{\top}), \tag{2}$$

where $\bar{h}_r$ is L2-normalized. The product $\pi(r;\mathcal{V})M_i(\mathcal{V})$ transfers class uncertainty into the shared coordinates. In parallel, the specialist predicts visual evidence $e_i(r) \in \mathbb{R}^J$. We combine them as

$$q_i^{\mathrm{attr}}(r) = \mathrm{Norm}_1([\pi(r;\mathcal{V})M_i(\mathcal{V})] \odot \sigma(e_i(r))) \in \mathbb{R}^J. \tag{3}$$

The two factors play complementary roles. The term $\pi(r;\mathcal{V})M_i(\mathcal{V})$ transfers the current class hypotheses into the shared coordinate space, whereas $\sigma(e_i(r))$ provides region-specific visual support over the same coordinates. Their element-wise interaction therefore allows the same queried class set to induce different adaptation signals across regions, while keeping the representation aligned to a common semantic basis. Thus the condition varies with both image and vocabulary while remaining J-dimensional. The class prior is detached from specialist optimization, while the frozen base remains the fixed recognition reference. Appendix A.2 gives normalization and calibration details.

3.5 Parameter level: composing the residual operator

A coefficient network controls $K = 4$ residual bases:

$$\rho_i(r) = \mathrm{Entmax}_{1.5}\big(G_i(q_i^{\mathrm{attr}}(r))/\tau_r\big), \qquad \widetilde{\rho}_i(r) = \alpha_i g_i \left(\rho_i(r) - \tfrac{1}{K}\mathbf{1}\right). \tag{4}$$

At layer ℓ,

$$\Delta_{i,\ell}^{\mathrm{AR}}(H,r) = \sum_{k=1}^{K} \widetilde{\rho}_{i,k}(r)\Delta_{i,\ell,k}(H), \tag{5}$$

where each $\Delta_{i,\ell,k}$ is low-rank (Hu et al., 2022). We use $\alpha_i = 1$, $g_i = 0.25$, internal rank 64, and final-projection rank 16. Centering makes the uniform mixture a zero-residual reference; non-uniform coefficients determine the update. All conditioner controls share the same anchor/capacity path. These shared components are initialized from the frozen base, optimized during the first increment, and frozen thereafter; their provenance and parameter accounting are detailed in Appendices A and D.5.

Changing the runtime vocabulary affects both the class–attribute relations $M_i(\mathcal{V})$ and the regional class evidence $\pi(r;\mathcal{V})$, and therefore can change $q_i^{\mathrm{attr}}(r)$. However, the resulting condition remains in $\mathbb{R}^J$, so the coefficient network, residual-bank size, and residual ranks do not scale with $|\mathcal{V}|$. Vocabulary changes are thus expressed through the class-dependent computation while the learned downstream adaptation capacity remains fixed.

3.6 Continual acquisition and deployment

Acquisition. For $t > 1$, E_t is initialized from E_{t-1} and trained only on the current domain; the base, shared fixed components, and earlier specialists remain frozen. The current specialist minimizes

$$\mathcal{L}_t = \mathcal{L}_{\mathrm{seg}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{rel}} + \lambda_{\mathrm{safe}}\mathcal{L}_{\mathrm{safe}} + \lambda_{\mathrm{repair}}\mathcal{L}_{\mathrm{repair}} + \lambda_{\mathrm{MI}}\mathcal{L}_{\mathrm{MI}}, \tag{6}$$

where relation supervision aligns visual evidence with the class–coordinate relation, safety/repair terms preserve or improve margins relative to a frozen reference path, and $\mathcal{L}_{\mathrm{MI}}$ encourages decisive but batch-diverse basis use (Appendix A.1). Since completed modules are frozen, for fixed preprocessing, image, and vocabulary,

$$f_i^{(t)}(x,\mathcal{V}) = f_i^{(i)}(x,\mathcal{V}), \qquad i < t, \tag{7}$$

up to numerical precision. Bank-level predictions can nevertheless change when retrieval and arbitration are recomputed over a larger specialist set.

Selection and arbitration without explicit task IDs. Each specialist stores frozen visual statistics (μ_i, v_i) and a native vocabulary reference C_i. Let $\bar{h}(x)$ denote the pooled frozen-base image representation. Retrieval combines a variance-normalized visual score with symmetric vocabulary coverage $c_i(\mathcal{V})$:

$$s_i^{\text{vis}}(x) = -\frac{1}{d}\sum_{a=1}^{d}\frac{(\bar{h}(x)_a - \mu_{i,a})^2}{\max(v_{i,a}, \varepsilon)},$$
$$s_i^{\text{retr}} = \text{Calib}_{\text{vis}}(s_i^{\text{vis}}) + \beta\,\text{Calib}_{\text{vocab}}(c_i(\mathcal{V})), \qquad i^* = \arg\max_i s_i^{\text{retr}}. \tag{8}$$

Only E_{i^*} is executed. Let $\ell_0(r)$ and $\ell_{i^*}(r)$ be base and specialist logits and let $m(\ell)$ denote the gap between the two largest softmax probabilities. Region utility is

$$a_{i^*}(r) = s_{i^*}^{\text{retr}} + \lambda_g[m(\ell_{i^*}(r)) - m(\ell_0(r))]. \tag{9}$$

A vocabulary gate compares the best specialist coverage $\max_j c_j(\mathcal{V})$ with the corresponding frozen-base coverage $c_0(\mathcal{V})$. The final regional prediction is

$$\ell(r) = \begin{cases} \ell_{i^*}(r), & \max_j c_j(\mathcal{V}) \geq c_0(\mathcal{V}) \text{ and } a_{i^*}(r) > 0, \\ \ell_0(r), & \text{otherwise.} \end{cases} \tag{10}$$

Thus retrieval decides *which* specialist is relevant, while discrete arbitration decides *where* specialization is applied and preserves exact local fallback to the frozen base. Appendix B gives the symmetric coverage definition, calibration, and thresholds.

4 Experiments

4.1 Experimental setup

Protocol and evaluation. Starting from a frozen Mask-Adapter-style COCO model (Lin et al., 2014; Li et al., 2025), we sequentially train specialists on A-150 (Zhou et al., 2017), Cityscapes (Cordts et al., 2016), and Mapillary Vistas (Neuhold et al., 2017) without replay, freezing the base and completed specialists. At inference, no explicit task identifier is provided; retrieval is followed by region-wise arbitration with optional fallback to the base, while forced-specialist activation is reported as a diagnostic. Retention is evaluated on COCO, PC-59 (Mottaghi et al., 2014), and PAS-20 (Everingham et al., 2010), with broader-taxonomy transfer on A-847 and PC-459. Training and order-swap details are provided in Appendices A.1.

Baselines. Table 1 compares VIRA with the frozen source, Mask-Adapter (Li et al., 2025), LoRA (Hu et al., 2022), adapter-only, EWC (Kirkpatrick et al., 2017), LwF (Li & Hoiem, 2016), ECLIPSE (Kim et al., 2024), and ConOVS (Hwang et al., 2025). This is a system-level comparison; Section 4.3 instead uses matched specialist and deployment settings to isolate the conditioning interface. Protocol details are in Appendix F.2.

Table 1: Final continual-system evaluation from a common COCO initialization, without explicit task IDs at inference. These endpoints are used as the deployment reference below.

Method	COCO		A-150		Cityscapes		Mapillary		PC-459	PAS-20	A-847	PC-59
	mIoU ↑	PQ ↑	mIoU ↑	PQ ↑	mIoU ↑	PQ ↑	mIoU ↑	PQ ↑	mIoU ↑	mIoU ↑	mIoU ↑	mIoU ↑
Frozen Source	64.31	52.37	34.12	25.48	52.87	38.02	25.43	11.90	18.63	96.02	13.42	59.17
Mask-Adapter	50.12	44.10	20.45	15.20	53.21	38.50	24.12	11.80	14.02	90.12	10.50	44.20
LoRA-only	39.85	35.10	10.92	8.10	35.42	25.50	14.80	7.20	15.90	60.45	8.12	24.30
Adapter-only	43.05	37.90	13.20	9.80	38.95	28.00	17.95	8.80	16.80	65.20	9.50	27.80
EWC	31.05	27.30	13.85	10.20	36.80	26.50	19.45	9.50	6.20	63.50	6.05	26.90
LwF	36.20	31.90	18.95	14.00	40.50	29.20	21.75	10.70	7.60	70.25	7.20	33.45
ECLIPSE	45.10	39.70	14.80	11.00	39.85	28.70	18.90	9.30	17.20	68.40	10.00	29.80
ConOVS	53.20	46.80	18.75	13.90	37.20	26.80	25.10	12.30	18.20	80.50	12.80	39.20
VIRA	**65.82**	**54.78**	**38.81**	**28.12**	**61.95**	**42.16**	**28.66**	**16.29**	**19.95**	**96.07**	**15.28**	**59.89**

4.2 Main continual results

Table 1 evaluates the continual system without explicit task IDs. VIRA improves three incremental domains over the frozen source. It also improves A-847 and PC-459, indicating that the benefit is not confined to specialist-native taxonomies. Among the systems, VIRA is the only method that improves all three incremental domains while maintaining or improving retention performance on COCO, PAS-20, and PC-59.

wrapfig

Figure 2 summarizes the cross-dataset trajectory. Incremental-domain gains coexist with non-degraded retained-source endpoints because arbitration can fall back to the frozen base. Since the base is frozen, retained-source gains cannot come from backward transfer; they arise from a region-wise mix of frozen-base fallback and selective specialist updates. Fallback ensures stability, and arbitration accepts specialists only where locally beneficial. Small gains over the frozen source therefore mean fallback preserves the base path, not that the final prediction is simply the frozen-source prediction.

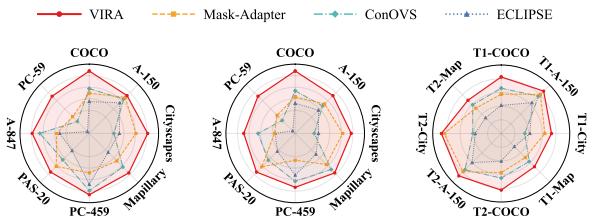

Figure 2: Cross-dataset continual performance profiles. The first two panels summarize mIoU after T1 and T2, and the third summarizes PQ.

4.3 Conditioning design controls

Table 2 asks whether the gains arise simply from exposing the conditioner to regional or query information. Under matched specialist capacity and deployment, VIRA consistently outperforms Region-only, Query-text-only, and Region+query-text conditioning. This supports the value of class-structured semantic coordinates beyond generic query access. Relative to the strongest generic Region+query-text control, VIRA gains +0.67/+0.63 macro mIoU/PQ, and exceeds Region-only by +1.41/+1.34. All controls share the same control dimension, residual bases, optimization budget, retrieval, arbitration, and frozen-base fallback; implementation details are provided in Appendix A.

Table 2: **Equal-capacity conditioner controls under deployment without explicit task IDs.** Results are means over three training runs; sample SDs are reported in Appendix F.1.

Conditioner	Params	A-150		City		Map		Macro	
		mIoU↑	PQ↑	mIoU↑	PQ↑	mIoU↑	PQ↑	mIoU↑	PQ↑
Region only	6.448M	37.52	26.88	60.55	40.86	27.11	14.82	41.73	27.52
Query text only	6.447M	37.96	27.41	61.27	41.49	27.42	15.18	42.22	28.03
Region + query text	6.448M	38.18	27.55	61.44	41.66	27.79	15.47	42.47	28.23
VIRA w/o $\mathcal{L}_{\text{rel}}$	6.447M	38.37	27.81	61.70	41.94	28.15	15.82	42.74	28.52
VIRA	6.447M	**38.81**	**28.12**	**61.95**	**42.16**	**28.66**	**16.29**	**43.14**	**28.86**

Table 3 further decomposes the structured conditioner. Removing the relation branch, visual evidence, dynamic residual composition, or relation supervision consistently weakens performance. The largest macro drops arise from removing dynamic composition or relation supervision, while removing either the relation or visual branch also consistently degrades performance. The rows are diagnostic rather than additive: removing the relation branch also removes its associated supervision, whereas the w/o-$\mathcal{L}_{\text{rel}}$ row keeps the relation parameterization. The dynamic-composition ablation retrains the model and is therefore distinct from the post-training coefficient interventions below.

4.4 Semantic role of attribute coordinates

We next ask whether the benefit comes from semantic attribute structure rather than a generic text-derived coordinate basis or target relevance. Automatic attribute coordinates consistently outperform relevance-matched non-attribute concepts across all three incremental domains, while random and

Table 3: **Component ablation of VIRA under deployment without explicit task IDs.** Results are means over three training runs and follow the same protocol as Table 2.

Relation	Visual	Dynamic	$\mathcal{L}_{\text{rel}}$	A-150		City		Map		Macro	
				mIoU↑	PQ↑	mIoU↑	PQ↑	mIoU↑	PQ↑	mIoU↑	PQ↑
×	✓	✓	–	38.50	27.89	61.63	41.90	28.35	16.01	42.83	28.60
✓	×	✓	✓	38.61	27.98	61.72	41.99	28.44	16.10	42.92	28.69
✓	✓	×	✓	38.44	27.82	61.55	41.83	28.26	15.94	42.75	28.53
✓	✓	✓	×	38.37	27.81	61.70	41.94	28.15	15.82	42.74	28.52
✓	✓	✓	✓	**38.81**	**28.12**	**61.95**	**42.16**	**28.66**	**16.29**	**43.14**	**28.86**

low-relevance controls are weaker (Table 4). This pattern supports an attribute-specific inductive bias beyond generic text structure and target relevance. The progression from random non-attribute to relevance-matched non-attribute and then to attribute coordinates separates the contribution of target relevance from that of attribute-specific structure. Construction details and matching statistics are provided in Appendix E.

Table 4: **Cross-domain coordinate-inventory controls.** Results are mean mIoU and PQ; variability across training runs or randomized inventory constructions is reported in the Appendix.

Inventory	A-150		City		Map		Macro	
	mIoU↑	PQ↑	mIoU↑	PQ↑	mIoU↑	PQ↑	mIoU↑	PQ↑
Curated Attribute 52	**38.81**	**28.12**	**61.95**	**42.16**	**28.66**	**16.29**	**43.14**	**28.86**
Relevance-matched Non-Attribute 52	38.21	27.54	61.42	41.64	28.12	15.76	42.58	28.31
Random Non-Attribute 52	37.43	26.73	60.85	41.02	27.55	15.20	41.94	27.65
LowRel 52	36.42	25.82	60.21	40.45	26.95	14.62	41.19	26.96
Gaussian 52	35.82	25.24	59.78	40.02	26.48	14.15	40.69	26.47

We then intervene on a trained model to test whether coordinate identity is actually used. On Cityscapes and Mapillary, relation uniformization and coordinate-column permutation cause much larger degradation than row permutation or within-image coefficient permutation (Figure 3). Column permutation largely preserves relation mass and sparsity while breaking the correspondence between relation entries and semantic coordinates. Its effect is therefore consistent with the relation and visual branches relying on a shared, aligned coordinate system. Figure 4 traces how that support is converted into region-specific residual updates.

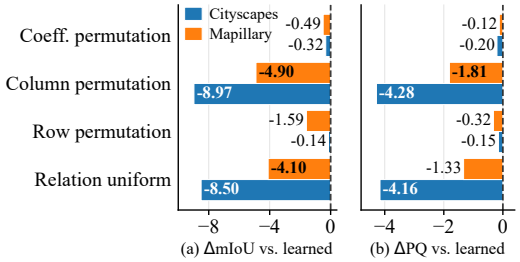

Figure 3: **Inference-time relation interventions.** Changes in mIoU and PQ relative to the learned model.

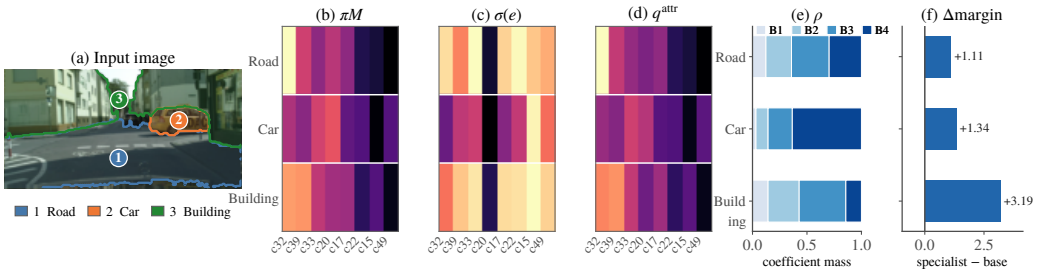

Figure 4: Region-level mechanism trace for road, car, and building regions through relation response, visual evidence, q^{attr}, residual coefficients ρ, and specialist-minus-base margin change.

Within the same specialist, road, car, and building regions activate different coordinate support, residual mixtures, and specialist-minus-base margin changes (Figure 4), illustrating how class hypotheses and local visual evidence meet in a shared semantic space before determining the residual operator. We then isolate the effect of runtime vocabulary by changing only the conditioning query while fixing the image, selected specialist, output vocabulary, retrieval result, evaluator, and arbitration. Here, **Subset-75** retains 75% of the native query set, whereas **Union-234** uses a fixed 234-class union vocabulary. Figure 5 shows that these perturbations propagate from q^{attr} through the residual coefficients and residual update to the final logits; Appendix Table 11 reports the corresponding endpoint effects.

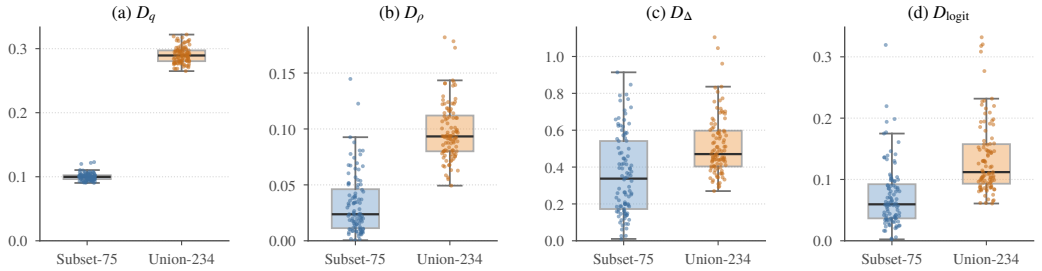

Figure 5: Signal propagation under conditioning-vocabulary interventions. Deviations from native conditioning are shown for q^{attr}, residual coefficients ρ, residual update Δ^{AR}, and output logits.

Both interventions affect every measured stage of the pathway. Union-234 produces larger changes in D_q, D_ρ, and D_{logit}, while D_Δ overlaps more. Subset-75 produces smaller but consistent shifts through the same pathway. This shows that runtime-vocabulary changes remain expressed through residual composition and are observable at the output.

Table 5: Post-training coefficient sensitivity under forced matched-specialist activation. "Learned" denotes the unperturbed mIoU under the same forced-specialist setting; the remaining columns report ΔmIoU under inference-time interventions to the learned coefficient patterns.

Dataset	Learned	Fixed mean	Within-image perm.	Across-image perm.	Uniform
A-150	38.89	−0.02	−0.03	0.00	−0.08
Cityscapes	62.22	−0.16	−0.32	−2.48	−8.11
Mapillary	29.69	−0.13	−0.49	−0.57	−4.07

Coefficient sensitivity is domain dependent. A-150 is nearly unchanged by the post-training coefficient interventions, whereas Cityscapes and Mapillary degrade substantially when the learned mixture is replaced by the centered uniform reference. Cityscapes is also sensitive to across-image coefficient reassignment. These results show that the importance of the precise mixture depends on the adapted domain. They should not be conflated with Table 3: that table removes dynamic composition during retraining, whereas Table 5 perturbs coefficients after the specialist has already been learned.

4.5 Deployment without explicit task IDs

Table 6: Prefix-bank trajectory without explicit task IDs; FB is frozen-base fallback.

Bank	Eval.	mIoU	PQ	FB (%)
T1	A-150	38.17	27.83	41.03
T2	A-150	37.65	27.45	67.78
	City	61.40	41.93	19.00
T3	A-150	38.81	28.12	28.56
	City	61.95	42.16	4.51
	Map	28.66	16.29	11.73

FB is the fraction of regions assigned to the frozen path. Historical specialist parameters are immutable, but deployed scores on earlier incremental domains can still move as the candidate bank grows because retrieval and arbitration are recomputed over a larger set of specialists. This separates parameter-level retention of completed modules from bank-level deployment behavior. Selector variants and the gap to forced activation are analyzed in Appendix F.3.

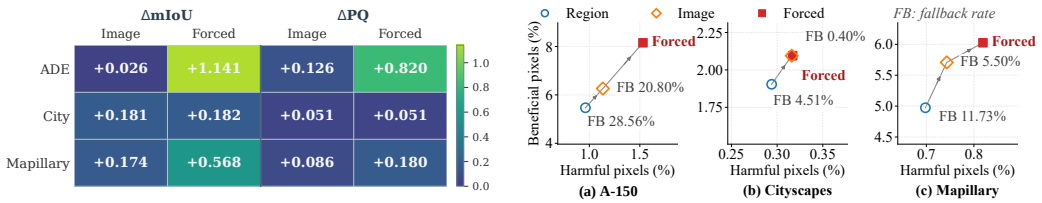

Figure 6: Final-bank arbitration effects relative to forced specialist activation, with corresponding harmful/beneficial-pixel and fallback rates.

Figure 6 shows that stronger specialist activation recovers additional incremental-domain performance but increases harmful pixel changes and reduces frozen-base fallback. Region-wise arbitration therefore trades some specialist gain for fewer harmful local changes while retaining the frozen path. We additionally assess whether dataset-native vocabularies act as implicit task cues through shared-vocabulary and visual-only routing diagnostics in Appendix B.1.

Qualitative behavior. Figure 7 compares two A-150 examples with ConOVS. VIRA produces more coherent semantic regions, suppresses isolated predictions, and better preserves object boundaries in both scenes. This behavior is consistent with region-conditioned residual adaptation, where local class priors and visual attribute evidence can induce different updates across image regions.

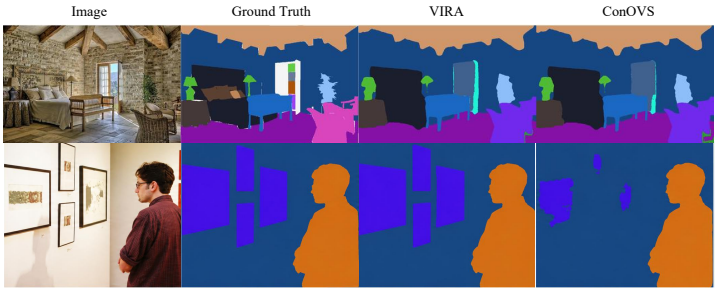

Figure 7: Illustrative A-150 cases. VIRA produces more coherent semantic-region predictions than ConOVS.

4.6 Transfer Scope

To test backbone transfer, we transplant the same interface to MAFT (Jiao et al., 2023). VIRA consistently outperforms the capacity-matched Region-only control across all three incremental domains while keeping PAS-20 essentially unchanged relative to the frozen source, supporting transfer beyond the Mask-Adapter-style backbone.

Table 7: Transfer to the MAFT OVS architecture under deployment without explicit task IDs.

Method	Params/inc.	A-150		City		Map		Macro		PAS-20
		mIoU↑	PQ↑	mIoU↑	PQ↑	mIoU↑	PQ↑	mIoU↑	PQ↑	
Region only	5.21M	36.84	26.42	58.74	40.02	26.72	14.72	40.77	27.05	95.58
VIRA	5.21M	**37.84**	**27.35**	**60.18**	**41.02**	**27.68**	**15.55**	**41.90**	**27.97**	**96.04**

5 Conclusion

We introduced VIRA, which maps open-vocabulary queries and regional evidence into shared semantic attribute coordinates that control low-rank residual adaptation. Class–attribute relations preserve query structure, visual evidence grounds it locally, and frozen specialists are deployed through retrieval and region-wise arbitration. Matched controls and non-attribute comparisons show gains beyond generic query access and target relevance, while interventions verify that the shared coordinates participate in adaptation. Routing diagnostics further show that deployment does not rely solely on dataset-native vocabulary cues. These results support semantic attributes as a reusable control space for continual OVS while retaining access to the frozen-source path.

REFERENCES

Tilemachos Aravanis, Vladan Stojnić, Bill Psomas, Nikos Komodakis, and Giorgos Tolias. Retrieve and segment: Are a few examples enough to bridge the supervision gap in open-vocabulary segmentation? In *CVPR*, pp. 27621–27632, 2026.

María A. Bravo, Sudhanshu Mittal, Simon Ging, and Thomas Brox. Open-vocabulary attribute detection. In *CVPR*, pp. 7041–7050, 2023.

Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulò, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *CVPR*, pp. 9233–9242, 2020.

Fabio Cermelli, Matthieu Cord, and Arthur Douillard. CoMFormer: Continual learning in semantic and panoptic segmentation. In *CVPR*, pp. 3010–3020, 2023.

Keyan Chen, Xiaolong Jiang, Yao Hu, Xu Tang, Yan Gao, Jianqi Chen, and Weidi Xie. OvarNet: Towards open-vocabulary object attribute recognition. In *CVPR*, pp. 23518–23527, 2023.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, pp. 4113–4123, 2024.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pp. 3213–3223, 2016.

Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. PLOP: Learning without forgetting for continual semantic segmentation. In *CVPR*, pp. 4040–4050, 2021.

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

Yizheng Gong, Siyue Yu, Xiaoyang Wang, and Jimin Xiao. Continual segmentation with disentangled objectness learning and class recognition. In *CVPR*, pp. 3848–3857, 2024.

Zhongrui Gui, Shuyang Sun, Runjia Li, Jianhao Yuan, Zhaochong An, Karsten Roth, Ameya Prabhu, and Philip Torr. knn-clip: Retrieval enables training-free segmentation on continually expanding large vocabularies. *Transactions on Machine Learning Research*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Dongjun Hwang, Yejin Kim, Minyoung Lee, Seong Joon Oh, and Junsuk Choe. Ovs meets continual learning: Towards sustainable open-vocabulary segmentation. In *Advances in Neural Information Processing Systems*, 2025.

Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. In *NeurIPS*, 2014.

Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Beomyoung Kim, Joonsang Yu, and Sung Ju Hwang. ECLIPSE: Efficient continual learning in panoptic segmentation with visual prompt tuning. In *CVPR*, pp. 3346–3356, 2024. doi: 10.1109/CVPR52733.2024.00322.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Naomi Kombol, Ivan Martinović, Siniša Šegvić, and Giorgos Tolias. SPAR: Single-pass any-resolution ViT for open-vocabulary segmentation. In *CVPR*, pp. 27579–27589, 2026.

Guannan Lai, Da-Wei Zhou, Zhenguo Li, and Han-Jia Ye. The golden subspace: Where efficiency meets generalization in continual test-time adaptation. In *CVPR*, pp. 3866–3875, 2026.

Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, and Xinggang Wang. Mask-adapter: The devil is in the masks for open-vocabulary segmentation. In *CVPR*, pp. 14998–15008, 2025.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *ECCV*, pp. 614–629, 2016. doi: 10.1007/978-3-319-46493-0_37.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Zihan Lin, Zilei Wang, and Xu Wang. Towards continual universal segmentation. In *CVPR*, pp. 29417–29427, 2025.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, pp. 891–898, 2014.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, pp. 4990–4999, 2017.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In *Annual Meeting of the Association for Computational Linguistics*, pp. 1504–1519, 2019.

Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *CVPR*, pp. 13018–13028, 2021.

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

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. PACO: Parts and attributes of common objects. In *CVPR*, pp. 7141–7151, 2023.

Lin Sun, Jiale Cao, Jin Xie, Xiaoheng Jiang, and Yanwei Pang. CLIPer: Hierarchically improving spatial representation of CLIP for open-vocabulary semantic segmentation. In *ICCV*, pp. 23199–23209, 2025.

Zhengyang Wang, Tingliang Feng, Fan Lyu, Fanhua Shang, Wei Feng, and Liang Wan. Dual semantic guidance for open vocabulary semantic segmentation. In *CVPR*, pp. 20212–20222, 2025.

Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. SED: A simple encoder-decoder for open-vocabulary semantic segmentation. In *CVPR*, pp. 3426–3436, 2024.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, pp. 2945–2954, 2023.

Xiwei Xuan, Ziquan Deng, and Kwan-Liu Ma. ReME: A data-centric framework for training-free open-vocabulary segmentation. In *ICCV*, pp. 20954–20965, 2025.

Hongmei Yin, Tingliang Feng, Fan Lyu, Fanhua Shang, Hongying Liu, Wei Feng, and Liang Wan. Beyond background shift: Rethinking instance replay in continual semantic segmentation. In *CVPR*, pp. 9839–9848, 2025.

Ziyu Zhao, Xiaoguang Li, Lingjia Shi, Nasrin Imanpour, and Song Wang. DPSeg: Dual-prompt cost volume learning for open-vocabulary semantic segmentation. In *CVPR*, pp. 25346–25356, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, pp. 633–641, 2017.

Ethics Statement

This work studies model adaptation and evaluation on established segmentation benchmarks and introduces no new personally identifiable data. As with other open-vocabulary perception systems, deployment may inherit biases from vision–language pretraining and may fail under ambiguous, culturally specific, or safety-critical vocabularies. In our evaluations, the frozen-base fallback reduces negative specialization changes and retains access to the base prediction path on retention domains. High-stakes uses warrant application-specific validation and human oversight.

Reproducibility Statement

Sections 3 and 4 specify the model, continual sequence, and evaluation protocol. The appendices provide the training objective, implementation and parameter accounting, retrieval/arbitration equations, coordinate-inventory construction, matched controls, and additional diagnostics. Tables 2 and 3 report means over three training runs; the corresponding sample standard deviations are reported in Appendix F.1. Fixed coordinate inventories report training-run variability where repeated, while construction-randomized inventories report construction-seed variability with the training seed fixed as stated.

Statement on the Use of Generative AI

Generative AI tools were used to assist with manuscript drafting, editing, and improving clarity and organization, and to provide feedback on the presentation and analysis of experimental controls and results. All experiments, numerical results, and final methodological decisions were produced or verified by the authors. Generative AI tools were not used to generate experimental data or reported results. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final content of this work.

A Implementation and Training Details

Frozen regional representation. For projected dense features $F_{\theta_0}(x) \in \mathbb{R}^{H\times W\times d}$ and region mask m_r,

$$h_r = \frac{\sum_p m_r(p) F_{\theta_0}(x)_p}{\sum_p m_r(p) + \epsilon}. \tag{11}$$

The base weights, projected feature path, and text embeddings remain frozen during specialist optimization. The implementation uses $d = 768$; a different base requires a compatible projection and initialization.

Class–attribute relation prior and correction. The compact class–attribute relation equation in the main text expands as

$$Z_{\mathcal{V}}(I + P_i Q_i^\top)A^\top = Z_{\mathcal{V}}A^\top + (Z_{\mathcal{V}}P_i)(AQ_i)^\top. \tag{12}$$

The shared inventory defines J coordinate names and embeddings, while each specialist learns its own relation correction and coefficient map.

Visual evidence and support normalization. With learned vectors $s_i, b_i \in \mathbb{R}^J$, the implemented evidence is

$$d_i = \exp(\mathrm{clip}(s_i, -\log 2, \log 2)), \tag{13}$$

$$e_i(r) = \mathrm{clip}\big(\mathrm{LN}(h_r)A^\top \odot d_i + b_i, -3, 3\big), \tag{14}$$

$$u_i(r) = \pi(r;\mathcal{V})M_i(\mathcal{V}), \tag{15}$$

$$q_i^{\mathrm{attr}}(r) = \frac{u_i(r)\odot\sigma(e_i(r))}{\max(\sum_a u_{i,a}(r)\sigma(e_{i,a}(r)), 10^{-8})}. \tag{16}$$

The detached prior uses $\kappa = 20$; relation normalization uses $\tau_a = 0.07$ and acts across attributes. Each query row is a distribution over the coordinate inventory. The resulting support vector is normalized for conditioning, not fitted as a probabilistically calibrated attribute posterior.

Residual parameterization. At adapted layer ℓ, basis k has

$$\Delta_{i,\ell,k}(H) = \frac{\gamma_\ell}{r_\ell} B_{i,\ell,k} A_{i,\ell,k} H. \tag{17}$$

Here $A_{i,\ell,k}$ is a low-rank factor, distinct from the shared attribute-text matrix A. There are four bases with internal rank 64 and final rank 16. The coefficient network has hidden width 128 and output temperature $\tau_r = 0.5$. The centered coefficients use $\alpha_i = 1$ and $g_i = 0.25$.

Full specialist path. In schematic layer notation,

$$H_{\ell+1}^{(i)} = f_\ell^{\mathrm{base}}(H_\ell) + \Delta_\ell^{\mathrm{anchor}}(H_\ell) + \Delta_\ell^{\mathrm{cap}}(H_\ell) + \Delta_{i,\ell}^{\mathrm{AR}}(H_\ell, r). \tag{18}$$

The anchor and two-component uniformly weighted capacity path are shared across specialists. They are initialized from the frozen base, optimized during the first increment, and frozen before later specialists are acquired. Uniform residual coefficients remove Δ^{AR} while retaining those components. The separate original-base prediction does not include these specialist additions. The following appendices provide the training objective, supervision details, and alias handling.

Matched conditioner definitions. The controls in Table 2 change only how query and region information is represented before the shared coefficient network. Let the runtime query embeddings be $Z_{\mathcal{V}} = \{z_v\}_{v\in\mathcal{V}} \in \mathbb{R}^{|\mathcal{V}|\times d}$ with $d = 768$. Both generic query-aware controls first apply the same learned attention-pooling design. For each query class v,

$$s_v = w_p^\top \tanh(W_p z_v + b_p), \tag{19}$$

$$\alpha_v = \frac{\exp(s_v)}{\sum_{v'\in\mathcal{V}}\exp(s_{v'})}, \tag{20}$$

$$\bar{z}_{\mathcal{V}} = \sum_{v\in\mathcal{V}} \alpha_v z_v, \tag{21}$$

where $W_p \in \mathbb{R}^{h_p \times d}$, $w_p, b_p \in \mathbb{R}^{h_p}$, and $h_p = 128$. The pooled vector $\bar{z}_\mathcal{V}$ is permutation invariant but does not retain an explicit class-by-coordinate relation matrix. Each control then produces a $J = 52$ dimensional condition:

$$q^{\text{region}}(r) = \text{MLP}_\psi(h_r) \in \mathbb{R}^J, \tag{22}$$

$$q^{\text{text}}(\mathcal{V}) = \text{MLP}_\theta(\bar{z}_\mathcal{V}) \in \mathbb{R}^J, \tag{23}$$

$$q^{\text{region+text}}(r, \mathcal{V}) = \text{MLP}_\phi([h_r; \bar{z}_\mathcal{V}]) \in \mathbb{R}^J, \tag{24}$$

$$q_i^{\text{attr}}(r) = \text{Norm}_1([\pi(r; \mathcal{V}) M_i(\mathcal{V})] \odot \sigma(e_i(r))) \in \mathbb{R}^J. \tag{25}$$

For Query-text-only, the same $q^{\text{text}}(\mathcal{V})$ is broadcast to every region, so the control is vocabulary-aware but not region-adaptive. Region+query text receives both region and pooled-vocabulary information but, unlike VIRA, does not preserve class-specific structure through $M_i(\mathcal{V})$ or ground that structure coordinate-wise with visual evidence. All four J-dimensional conditions are passed to the same coefficient-network architecture G_i, which produces $K = 4$ Entmax coefficients and controls the same residual bases. The controls also share the same training budget, retrieval, arbitration, and frozen-base fallback. Projection widths are chosen to closely match specialist capacity. The exact counts are 6,447,527 (Region only), 6,447,444 (Query text only), 6,447,955 (Region + query text), 6,447,444 (VIRA without $\mathcal{L}_{\text{rel}}$), and 6,447,444 (VIRA), so the remaining differences are below 0.01%.

Control structure and scaling. All matched query-aware controls feed the same $J = 52$ downstream controller, so generic pooling already provides dimensional compatibility. The distinction in VIRA is instead the use of shared coordinates that retain class-specific query structure and admit region-level visual grounding. The number of learned specialist parameters is independent of $|\mathcal{V}|$, while relation evaluation and class scoring scale with query size. The experiments separately test generic query access, coordinate semantics and target relevance, vocabulary sensitivity, and continual deployment behavior.

A.1 Objective and Training Configuration

The current specialist is optimized with the objective

$$\mathcal{L}_t = \mathcal{L}_{\text{seg}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}} + \lambda_{\text{safe}}\mathcal{L}_{\text{safe}} + \lambda_{\text{repair}}\mathcal{L}_{\text{repair}} + \lambda_{\text{MI}}\mathcal{L}_{\text{MI}}. \tag{26}$$

Class supervision through attribute coordinates. Let $\mathcal{R}_t$ be regions with a valid class target y_r, obtained from ground-truth masks and the base training pipeline's IoU-matched predicted masks. The class–attribute relation classifier uses the visual evidence from Section 3.4:

$$\mathcal{L}_{\text{rel}} = \frac{1}{|\mathcal{R}_t|} \sum_{r \in \mathcal{R}_t} \text{CE}\big(\kappa e_t(r) M_t(\mathcal{V}_t)^\top, y_r\big)\,. \tag{27}$$

We use $\kappa = 20$ and $\lambda_{\text{rel}} = 0.01$. A minibatch with no valid relation targets contributes zero relation loss. Supervision is on classes, not annotated attribute values; attribute names provide text coordinates without requiring disentangled attribute predictions.

Reference-margin retention and repair. The training reference is the frozen anchor/capacity path with the relation-dependent residual disabled; it is distinct from the untouched deployment base. Let l_{ref} and l_t denote reference and student logits, and define the ground-truth margin

$$m(l, y) = l_y - \max_{c \neq y} l_c, \qquad m_{\text{ref}}(r) = m(l_{\text{ref}}(r), y_r), \quad m_t(r) = m(l_t(r), y_r). \tag{28}$$

Let $\mathcal{S}$ contain valid regions correctly predicted by the reference, and $\mathcal{U}$ the remaining valid regions. If $w_y > 0$ is the stored training-class weight and $c_{\text{ref}}(r) = \max_c \text{softmax}(l_{\text{ref}}(r))_c$, the implemented losses are

$$\mathcal{L}_{\text{safe}} = \frac{\sum_{r \in \mathcal{S}} w_{y_r} c_{\text{ref}}(r) [\eta \max(m_{\text{ref}}(r), 0) - m_t(r)]_+}{\max(\sum_{r \in \mathcal{S}} w_{y_r} c_{\text{ref}}(r), \epsilon)}, \tag{29}$$

$$\mathcal{L}_{\text{repair}} = \frac{\sum_{r \in \mathcal{U}} w_{y_r} [m_{\text{ref}}(r) + \delta - m_t(r)]_+}{\max(\sum_{r \in \mathcal{U}} w_{y_r}, \epsilon)}. \tag{30}$$

Reference logits and weights are detached, and empty sets produce zero loss. Deployment-time base–specialist arbitration provides a separate mechanism for limiting unfavorable specialization changes.

Coefficient utilization. Writing $\bar{\rho}_t$ for the minibatch mean coefficient vector,

$$\mathcal{L}_{\mathrm{MI}} = -H(\bar{\rho}_t) + \mathbb{E}_r[H(\rho_t(r))]. \tag{31}$$

This favors utilization across a minibatch and decisive coefficients within regions.

Configuration. The retained configuration has four residual bases, internal rank 64, final rank 16, coefficient-network hidden width 128, temperature 0.5, MI weight 0.01, fixed residual scale $\alpha = 1$, and centered gain $g = 0.25$. The shared capacity path has two components with gain 0.25. Safety/repair weights are 0.375/0.0625, with $\eta = 1$ and $\delta = 0.25$; separate balance and gate losses are disabled. The relation temperature is 0.07. Optimization uses four workers, global batch size 4, initial learning rate 10^{-4}, weight decay 0.05, and 500 learning-rate warm-up iterations. Each increment uses a 20k-update budget, with later-increment checkpoints retained at 5k intervals. Input mapping uses training size 1024 and test size 896 with training-time horizontal flips.

A.2 Interface Properties and Runtime Details

The identities below summarize the interface properties used in the method description; Appendix C gives statements and proofs for query-order equivariance, residual bounds, stability, and continual/fallback properties.

For a permutation matrix Π acting on the query classes, $Z_{\mathcal{V}'} = \Pi Z_{\mathcal{V}}$. Because relation normalization acts independently on each row,

$$M_i(\mathcal{V}') = \Pi M_i(\mathcal{V}), \qquad \pi(r; \mathcal{V}') = \pi(r; \mathcal{V})\Pi^\top. \tag{32}$$

Thus $\pi(r; \mathcal{V}')M_i(\mathcal{V}') = \pi(r; \mathcal{V})M_i(\mathcal{V})$, and attribute support and residual coefficients are invariant to a consistent class reordering. Class logits are correspondingly permuted, subject to identical text preprocessing and numerical/tie handling. Adding or removing classes can change the normalized prior and therefore the coefficients. This gives class-order equivariance; sensitivity to changed vocabulary content and aliases is evaluated separately through the conditioning-vocabulary and synonym controls.

Centering also gives

$$\Delta^{\mathrm{AR}}_{i,\ell} = \alpha_i g_i \sum_k \rho_{i,k}(r) \left(\Delta_{i,\ell,k} - \frac{1}{K} \sum_j \Delta_{i,\ell,j} \right). \tag{33}$$

Uniform coefficients cancel the centered residual. Uniform relation entries do not imply uniform coefficients, because visual support and the coefficient network remain active.

Alias ensembles and prediction heads. The main equations use one text embedding per query class for readability. In the runtime evaluator, a canonical category can contain several aliases. The within-specialist relation and frozen prior are evaluated on the expanded alias embeddings; canonical category classification takes the maximum alias score using the base model's convention. Consistent class reordering must preserve these groups. Reducing an ensemble to one alias therefore changes both the linguistic input and aggregation budget.

Arbitration selects the out-of-vocabulary branch logits. The original detector branch, geometric fusion for seen/unseen categories, void-class probability, mask predictions, and semantic/panoptic postprocessing are then applied identically to the selected and base-reference logits. The unchanged masks keep candidate regions aligned. For Union-234 evaluation, a fixed mapping projects query-space probabilities to native evaluator classes; this projection is downstream of retrieval and adaptation.

B Deployment Retrieval and Arbitration without Explicit Task IDs

Retrieval and arbitration equations. Each specialist stores a frozen visual mean and diagonal variance (μ_i, v_i) and native vocabulary embeddings C_i. Retrieval combines a variance-normalized

visual score and symmetric text coverage:

$$s_i^{\mathrm{vis}}(x) = -\frac{1}{d}\sum_{a=1}^{d}\frac{(\bar{h}(x)_a - \mu_{i,a})^2}{\max(v_{i,a}, \varepsilon)}, \tag{34}$$

$$c_i(\mathcal{V}) = \frac{1}{2}\left[\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\max_{c\in C_i}\cos(z_v, z_c) + \frac{1}{|C_i|}\sum_{c\in C_i}\max_{v\in\mathcal{V}}\cos(z_v, z_c)\right], \tag{35}$$

$$s_i^{\mathrm{retr}} = \mathrm{Calib}_{\mathrm{vis}}(s_i^{\mathrm{vis}}) + \beta\,\mathrm{Calib}_{\mathrm{vocab}}(c_i). \tag{36}$$

Here $\bar{h}(x)$ is the L2-normalized spatial mean of the projected dense features. Calibration uses fixed training feature caches. Validation selection retains top-$K = 1$; the bank remains searchable as specialists accumulate.

Let $l_0(r), l_i(r)$ be base and retrieved-specialist class logits, and let $m(l)$ be the difference between the largest two entries of $\mathrm{softmax}(l)$. The selected configuration uses

$$a_i(r) = s_i^{\mathrm{retr}} + \lambda_g[m(l_i(r)) - m(l_0(r))], \qquad a_0(r) = 0, \tag{37}$$

with $\lambda_g = 4$ and no outer attribute-consistency term. The confidence difference is a proxy, not an estimate of ground-truth improvement. A vocabulary gate compares $\max_j c_j(\mathcal{V})$ with the base vocabulary coverage $c_0(\mathcal{V})$. With zero acceptance and decision margins, the deployed top-1 rule is

$$l(r) = \begin{cases} l_i(r), & \max_j c_j(\mathcal{V}) \geq c_0(\mathcal{V}) \text{ and } a_i(r) > 0, \\ l_0(r), & \text{otherwise.} \end{cases} \tag{38}$$

Exact score ties are resolved deterministically. Indexed logit selection makes base fallback possible. Output selection is discrete rather than a soft mixture, and the within-specialist class–attribute-conditioned residual remains active whenever a specialist is evaluated.

Calibration and diagnostics. The visual score uses diagonal population variances with floor 10^{-4}. Each score type is centered and scaled using fixed training caches; the selected vocabulary weight is $\beta = 1$. The vocabulary score is the average of query-to-reference and reference-to-query nearest-neighbor cosine coverage. Base acceptance compares the maximum specialist vocabulary score with the base vocabulary score, using margin zero. For the native retention vocabularies, this vocabulary-level decision can reject the entire specialist bank before local selection.

The retained top-1 region score is given in equation 37. The deployed rule uses no trained outer classifier, supervised utility-matching objective, or probability interpolation. The optional outer attribute term and the separate nonnegative-gain gate are disabled. Fallback statistics therefore count regions assigned to the frozen base by the discrete arbitration rule.

Top-K validation macro mIoU is 42.13/42.14/42.14 for $K = 1/2/3$, respectively. We therefore retain top-1 as the smallest candidate count within 0.2 mIoU of the best validation score.

The nearly identical top-K validation scores suggest that retrieval breadth is not a major factor for this bank: top-1 captures nearly all of the observed validation benefit. The selector variants expose a different trade-off. Vocabulary-nearest selection gives higher incremental performance but lower source retention, whereas global fallback preserves retention while rejecting more incremental cases. Region-level arbitration recovers much of the specialist gain while keeping the frozen path available locally. This gives the deployed division of labor: retrieval chooses a candidate specialist, while arbitration decides where that candidate is used.

B.1 Dataset-Vocabulary Cue Diagnostic

The deployed selector may use the runtime vocabulary because it is part of the OVS input, but benchmark-native query sets can also reveal dataset identity. We therefore perform inference-only routing diagnostics that remove this cue without retraining any specialist. These diagnostics test whether routing depends on dataset-native vocabulary identity; they do not test invariance of the within-specialist conditioner to vocabulary content. In all non-oracle variants, the conditioner and output vocabulary remain native to the evaluation dataset. Shared-vocab retrieval replaces the retrieval vocabulary for every image with the same fixed Union-234 query set; Shared-vocab, no gate

additionally disables the vocabulary-level base-acceptance gate; and Visual-only retrieval sets $\beta = 0$ and disables the vocabulary gate. Specialist accuracy uses ground-truth dataset identity only for offline scoring and is never provided to the model. The oracle variant supplies the correct specialist for analysis and bypasses retrieval and base arbitration.

Frozen visual-prototype diagnostics give 99.22%, 99.22%, and 97.27% top-1 specialist retrieval on A-150, Cityscapes, and Mapillary in the validation analysis. On the retention domains, the vocabulary gate rejects the entire specialist bank for 21.09%, 16.02%, and 20.31% of COCO, PC-59, and PAS-20 images, respectively. These are bank-level rejection rates, not region-level frozen-base fallback rates. They indicate that retrieval and bank acceptance do not rely solely on dataset-native vocabulary cues, but they do not by themselves explain the retained-source endpoints. As described in the main text, retained-source behavior reflects a region-wise combination of exact frozen-base fallback and selective specialist updates; region-level fallback is controlled by the arbitration rule, not by the bank-level vocabulary gate alone.

Table 8: Routing diagnostic for dataset-vocabulary cues. All rows use native conditioner/output vocabularies. Specialist Acc. reports A-150/Cityscapes/Mapillary top-1 specialist accuracy in percent.

Routing setting	Retrieval vocab.	Gate	A-150	City	Map	Macro	Specialist Acc.
			mIoU / PQ				A / C / M (%)
Native deployed	Native	✓	38.81/28.12	61.95/42.16	28.66/16.29	43.14/28.86	99.2/99.2/97.3
Shared-vocab retrieval	Union-234	✓	38.65/28.01	61.72/41.98	28.41/16.10	42.93/28.70	98.8/98.9/96.5
Shared-vocab, no gate	Union-234	×	38.52/27.91	61.58/41.85	28.24/15.96	42.78/28.57	98.1/98.3/95.7
Visual-only retrieval	None ($\beta = 0$)	×	38.21/27.68	61.20/41.52	27.85/15.61	42.42/28.27	95.6/96.1/92.8
Oracle forced specialist	Oracle label	–	38.89/28.13	62.22/42.22	29.69/16.99	43.60/29.11	100/100/100

Table 9: Visual-only specialist-selection confusion matrix (%). Rows are ground-truth datasets and columns are selected specialists.

True \ selected	A-150	City	Map
A-150	95.6	2.1	2.3
Cityscapes	1.5	96.1	2.4
Mapillary	3.8	3.4	92.8

We also apply the same shared Union-234 retrieval vocabulary to the retention datasets while preserving their native output vocabularies. This removes dataset-native vocabulary identity from retention routing.

C Formal Properties of the VIRA Interface

This appendix records formal properties of the construction. They are algebraic consequences rather than learning or generalization guarantees; the corresponding empirical behaviors are examined through the intervention, vocabulary-control, and deployment experiments.

C.1 Parameter scaling and query-order symmetry

Let $Z_{\mathcal{V}} \in \mathbb{R}^{|\mathcal{V}| \times d}$, $A \in \mathbb{R}^{J \times d}$, and

$$M_i(\mathcal{V}) = \mathrm{Entmax}_{1.5}\big(Z_{\mathcal{V}}(I + P_i Q_i^\top)A^\top/\tau_a\big)\,.$$

The learned factors P_i, Q_i and downstream coefficient network have dimensions independent of $|\mathcal{V}|$; changing the runtime vocabulary changes the number of relation rows and class-scoring work, but not the learned specialist parameter count. The correction $P_i Q_i^\top$ has rank at most r_a.

For any class permutation matrix Π, setting $Z_{\mathcal{V}'} = \Pi Z_{\mathcal{V}}$ gives

$$M_i(\mathcal{V}') = \Pi M_i(\mathcal{V}), \qquad \pi(r;\mathcal{V}') = \pi(r;\mathcal{V})\Pi^\top,$$

and hence $\pi(r;\mathcal{V}')M_i(\mathcal{V}') = \pi(r;\mathcal{V})M_i(\mathcal{V})$. Therefore the regional attribute condition and residual coefficients are invariant to a consistent reordering of the query classes, while final class logits are equivariant to the same permutation. This does not imply invariance to adding or removing classes or changing aliases, because those operations change normalized class evidence; the conditioning-vocabulary interventions and synonym control test those cases separately.

C.2 Centered residual composition, rank, and stability

Because $\rho_i(r)$ lies on the probability simplex, the centered coefficients

$$\widetilde{\rho}_i(r) = \alpha_i g_i \left(\rho_i(r) - K^{-1}\mathbf{1}\right)$$

sum to zero. Thus

$$\Delta^{\mathrm{AR}}_{i,\ell}(H, r) = \alpha_i g_i \left[\sum_k \rho_{i,k}(r)\Delta_{i,\ell,k}(H) - \frac{1}{K}\sum_k \Delta_{i,\ell,k}(H)\right],$$

so uniform coefficients yield zero attribute-conditioned residual. If each basis operator has rank at most r_ℓ, then for fixed coefficients the composed linear residual has rank at most Kr_ℓ.

Let $D_{i,\ell}(H) = \max_{j,k} \|\Delta_{i,\ell,j}(H) - \Delta_{i,\ell,k}(H)\|$. For any two coefficient vectors ρ, ρ' on the simplex,

$$\|\Delta^{\mathrm{AR}}(\rho) - \Delta^{\mathrm{AR}}(\rho')\| \le \frac{|\alpha_i g_i| D_{i,\ell}(H)}{2}\|\rho - \rho'\|_1.$$

The bound follows by viewing the positive and negative parts of $\rho - \rho'$ as two convex combinations of the same basis responses. It is conditional on fixed basis responses and does not claim global Lipschitzness of the full nonlinear network. For the normalized attribute support $q(u, e) = \mathrm{Norm}_1(u \odot \sigma(e))$, clipping $e \in [-3, 3]^J$ keeps the normalization denominator positive and gives finite local sensitivity to changes in class-induced support and visual evidence. The main-text signal-propagation experiment measures the realized sensitivity to vocabulary changes.

C.3 Continual immutability, fallback, and bank scaling

Once specialist E_i is completed, its parameters and stored statistics are frozen. Consequently, for fixed preprocessing, image, vocabulary, and forced specialist identity, later training increments do not change its output except for numerical nondeterminism. Bank-level deployed predictions can nevertheless change because later specialists alter candidate retrieval and arbitration.

The discrete arbitration rule selects either specialist logits or unchanged base logits. If the vocabulary-level gate rejects the bank, or if every region chooses the base branch, the selected logits are exactly the frozen-base logits before common downstream postprocessing. This is the exact-fallback property used in the retention analysis.

With t increments, specialist storage grows linearly as tP_E for per-specialist parameter cost P_E, plus fixed shared components. Top-1 execution need not activate all stored specialists, but retrieval and storage still scale with bank size; T3 parameter and storage costs are reported in Appendix D.5.

C.4 Coordinate reparameterization perspective

For a simplified linear conditioner $c = Wq$, an invertible coordinate change $q' = Tq$ can be absorbed by $W' = WT^{-1}$ without changing c. This explains why multiple well-structured coordinate systems can sometimes be absorbed downstream. It is not an invariance theorem for the full VIRA pipeline: changing A also changes relation logits and visual evidence before entmax, sigmoid, and normalization. The inventory experiments therefore directly test how semantic structure and target relevance affect the nonlinear system.

D Additional Controls, Sensitivity, and Efficiency

D.1 Region-Level Base–Specialist Arbitration

In the global vocabulary-margin diagnostic, the specialist-vs-base acceptance switches occur at margins 1.83, 2.08, and 2.57 for A-150, Cityscapes, and Mapillary. Because this margin is applied to a vocabulary-level quantity that is constant for all images evaluated with the same query set, it acts as a dataset-level switch: margin 1.9 rejects A-150, 2.2 rejects A-150 and Cityscapes, and 2.6 rejects all three. It does not distinguish local positive and negative region-level changes; accordingly, the deployed global margin is zero and local base–specialist arbitration is retained.

This behavior exposes a granularity mismatch in vocabulary-only acceptance. Vocabulary compatibility can reject a mismatched bank at the dataset/query level, but once a specialist is accepted it does not distinguish positive and negative region-level changes within the same image. Region-level arbitration supplies that local decision and complements the vocabulary gate.

D.2 Panoptic Intervention Metrics

The SQ/RQ breakdown follows the same qualitative pattern as the main intervention results: on Cityscapes and Mapillary, relation uniformization and coordinate-column permutation reduce both segmentation and recognition/association quality, whereas relation-row and within-image coefficient permutations have much smaller effects. We therefore report the more compact mIoU/PQ intervention summary in the main text and omit redundant per-metric tables here.

D.3 Relation Ablations and Cross-Specialist Results

An additional A-150 intervention uses two random permutations per intervention. Average learned-minus-intervention differences are 0.08/0.05 for class–attribute relation row permutation, 0.17/0.02 for class–attribute relation column permutation, and $0.04/-0.02$ for region-coefficient permutation. These A-150 effects are small relative to the road-domain interventions.

Table 10: Cross-specialist performance under forced specialist activation. Each cell reports the absolute score followed by the change relative to the target-matched specialist in parentheses.

Metric	Target	Force A-150	Force City	Force Map
mIoU	A-150	**38.89** (0.00)	38.61 (-0.28)	38.52 (-0.37)
	City	54.10 (-8.12)	**62.22** (0.00)	56.13 (-6.09)
	Map	25.62 (-4.07)	26.57 (-3.12)	**29.69** (0.00)
PQ	A-150	**28.13** (0.00)	28.08 (-0.05)	28.29 (+0.16)
	City	38.48 (-3.74)	**42.22** (0.00)	39.87 (-2.35)
	Map	15.55 (-1.44)	15.78 (-1.21)	**16.99** (0.00)

Forced cross-specialist evaluation shows domain-specific behavior in the learned residuals. Cityscapes and Mapillary incur larger semantic drops under a non-matched road specialist, while A-150 changes less in mIoU and slightly in PQ. The largest mismatched-activation drops occur on the road domains; A-150 shows weaker specialist dependence, including a slight PQ increase under the Mapillary specialist.

D.4 Conditioning-Vocabulary Controls

These interventions probe whether runtime query changes enter the learned within-specialist adaptation pathway while keeping deployment decisions fixed; they are sensitivity tests rather than claims of invariance to vocabulary changes.

Signal-deviation metrics. For the diagnostics in Figure 5, let $\mathcal{V}_0$ denote the native conditioning vocabulary, $\mathcal{V}'$ a perturbed conditioning vocabulary, and $\mathcal{R}$ the valid regions of an image. With the image, specialist, output vocabulary, retrieval result, evaluator, and arbitration fixed, we compute

$$D_q = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \left\| q^{\mathrm{attr}}(\mathcal{V}', r) - q^{\mathrm{attr}}(\mathcal{V}_0, r) \right\|_1, \tag{39}$$

$$D_\rho = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \left\| \rho(\mathcal{V}', r) - \rho(\mathcal{V}_0, r) \right\|_1, \tag{40}$$

$$D_\Delta = \frac{\left\| \Delta^{\mathrm{AR}}(\mathcal{V}') - \Delta^{\mathrm{AR}}(\mathcal{V}_0) \right\|_2}{\max(\|H_{\mathrm{ref}}\|_2, 10^{-8})}, \tag{41}$$

$$D_{\mathrm{logit}} = \sqrt{\frac{1}{|\mathcal{R}|C} \sum_{r \in \mathcal{R}} \sum_{c=1}^{C} \left(\ell_{r,c}(\mathcal{V}') - \ell_{r,c}(\mathcal{V}_0) \right)^2}. \tag{42}$$

Here H_{ref} is the native hidden-state reference and $\ell(\mathcal{V}) \in \mathbb{R}^{|\mathcal{R}|\times C}$ denotes region-level logits computed in the fixed native output space. Thus D_q and D_ρ are region-averaged ℓ_1 deviations, D_Δ is a hidden-state-normalized residual ℓ_2 deviation, and D_{logit} is the RMS logit shift. Each image contributes one scalar per metric; the native comparison $\mathcal{V}' = \mathcal{V}_0$ is zero by definition.

Table 11: Conditioner-only query interventions with image, specialist, output vocabulary, retrieval, evaluator, and arbitration held fixed. These absolute endpoints come from the dedicated query-intervention sweep and are compared within this table.

Conditioning query	City mIoU	City PQ	Map mIoU	Map PQ
Original query	**62.15**	**42.18**	**29.60**	**16.96**
Subset-75	61.82	41.94	29.23	16.74
Union-234	61.48	41.68	28.91	16.52

The conditioner-only intervention shows that changing the conditioning vocabulary while holding the specialist and output vocabulary fixed changes the road-domain endpoints, consistent with query semantics participating in adaptation. The synonym control, by contrast, keeps the VIRA-minus-base A-150 gain similar across the tested naming variants. These results establish controlled vocabulary sensitivity, not invariance to arbitrary query changes.

D.5 Efficiency Details

Table 12: Parameter and storage accounting. Percentages are relative to the 387.40M frozen base.

Component	Parameters	Relative
Per specialist	6.45M	1.66%
Shared anchor	6.41M	1.65%
Shared capacity	3.17M	0.82%
T3 cumulative additions	28.92M	7.47%
Top-1 active additions	16.03M	4.14%
Top-2 active additions	22.47M	5.80%

The shared anchor and capacity components are initialized from the frozen base, optimized during the first increment, and frozen thereafter. Thus 1.66% is marginal specialist storage; top-1 active and T3 cumulative additions are 4.14% and 7.47%, respectively.

E Coordinate-Inventory and Attribute-Specificity Study

This study asks whether the benefit of the shared coordinate representation comes from attribute structure rather than generic text-derived geometry or target relevance. Each variant changes the coordinates used by both the class–attribute relation and visual-evidence branches. The detailed panoptic decomposition uses A-150, while attribute-specificity controls are also evaluated on Cityscapes and Mapillary under the same matched protocol. For construction-randomized controls, only the inventory changes across construction seeds while the training seed is fixed unless stated otherwise.

Inventory construction. Here J counts coordinate columns, not a prescribed number of attributes assigned to each class.

- **Curated 52:** the default frozen text embeddings for 52 manually assembled color, material, geometry, appearance, and state attributes.
- **Relevance-matched Non-Attribute 52:** construct a non-attribute text-concept pool from object nouns, scene nouns, actions/events, and generic semantic concepts, while excluding color, material, shape, state, appearance, and other property terms. Select 52 concepts to match Automatic-52 in target-relevance distribution and coordinate geometry. Five construction seeds produce independent matched 52-concept inventories under these same relevance and geometry criteria, while the training seed remains 42.

- **Random Non-Attribute 52:** sample 52 concepts uniformly from the same non-attribute pool. Five construction seeds vary the sampled inventory while the training seed remains 42.
- **LowRel 52:** rank the same 620 VAW candidates by maximum cosine similarity to the target-domain class text embeddings, retain the 256 lowest-relevance candidates, and select 52 by farthest-point sampling. Construction seeds 20260912–20260916 vary the selected inventory while the training seed remains 42.
- **Gaussian 52:** draw 52 Gaussian vectors in $\mathbb{R}^{768}$, L2-normalize each, and freeze them. We use construction seeds 20260907–20260911 while holding the training seed fixed at 42.

The automatic prompt template differs from the curated prompts, and the 26/104 variants jointly vary semantic content and coordinate dimension. Random-VAW and LowRel retain the VAW text-attribute space and $J = 52$ while changing target relevance. The two non-attribute controls isolate attribute type from generic text-manifold membership and target alignment. For the matching statistics in Table 13, we summarize class-to-coordinate relevance as

$$r_{\mathrm{match}}(a) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \cos(z_v, z_a), \tag{43}$$

and additionally report relevance quartiles and mean pairwise coordinate cosine.

Table 13: Matching statistics for the attribute-specificity controls. Relevance uses Eq. (43).

Inventory	Mean relevance	Q1 / Q2 / Q3	Pairwise cosine mean
Automatic Attribute 52	0.214	0.16 / 0.21 / 0.27	0.19
Relevance-matched Non-Attribute 52	0.212	0.16 / 0.21 / 0.27	0.18
Random Non-Attribute 52	0.108	0.06 / 0.11 / 0.16	0.14

Table 14: Image-level paired-bootstrap differences for the A-150 attribute-specificity controls. Predictions are first averaged across the five construction-specific runs for randomized inventories; bootstrap resampling then pairs methods on the same evaluation images. Intervals therefore quantify image-level evaluation uncertainty after construction averaging, not training-seed uncertainty.

Comparison	ΔmIoU	95% CI	ΔPQ	95% CI
Curated Attr. − Matched Non-Attr.	+0.60	[0.39, 0.93]	+0.58	[0.32, 0.84]
Matched Non-Attr. − Random Non-Attr.	+0.78	[0.49, 1.07]	+0.81	[0.52, 1.10]
Random Non-Attr. − Random-VAW	−0.09	[−0.31, 0.13]	−0.07	[−0.29, 0.15]

F Additional Matched and Continual Experiments

This section provides supporting evidence for three roles separated in the main text: system-level continual comparison, matched analysis of the conditioner, and deployment behavior of the accumulated specialist bank. Tables 2 and 3 use three training runs; other inference-time mechanistic interventions use seed 42 unless stated otherwise. Table 1 remains a system-level comparison, whereas the matched tables isolate conditioner design under fixed specialist/deployment settings.

F.1 Run-to-run variability for matched controls

Tables 15 and 16 report the sample standard deviations corresponding to the three-run means shown in Tables 2 and 3. The main text omits these values for readability.

F.2 System-level baseline comparison protocol

Table 1 baseline protocol. All methods start from the same frozen COCO-pretrained Mask-Adapter-style open-vocabulary segmentation model. The continual stream is A-150 → Cityscapes → Mapillary Vistas, with no replay of historical images. Each method retains its native training objective and update rule: Mask-Adapter updates the model on each increment; LoRA-only and Adapter-only

Table 15: Sample SD over three training runs for the matched conditioner controls in Table 2. Each entry is mIoU/PQ. Exact specialist parameter counts are listed in Appendix A.

Conditioner	A-150	City	Map	Macro
Region only	0.13/0.17	0.11/0.15	0.16/0.19	0.09/0.12
Query text only	0.12/0.16	0.10/0.14	0.15/0.18	0.08/0.11
Region + query text	0.11/0.15	0.10/0.13	0.14/0.17	0.08/0.11
VIRA w/o $\mathcal{L}_{\mathrm{rel}}$	0.12/0.16	0.09/0.13	0.14/0.17	0.07/0.10
VIRA	0.10/0.14	0.08/0.12	0.13/0.16	0.07/0.10

Table 16: Sample SD over three training runs for the component ablations in Table 3. Each entry is mIoU/PQ.

Relation	Visual	Dynamic	$\mathcal{L}_{\mathrm{rel}}$	A-150	City	Map	Macro
×	✓	✓	–	0.12/0.16	0.10/0.14	0.15/0.18	0.08/0.11
✓	×	✓	✓	0.11/0.15	0.09/0.13	0.14/0.17	0.07/0.10
✓	✓	×	✓	0.13/0.17	0.11/0.15	0.16/0.19	0.09/0.12
✓	✓	✓	×	0.12/0.16	0.09/0.13	0.14/0.17	0.07/0.10
✓	✓	✓	✓	0.10/0.14	0.08/0.12	0.13/0.16	0.07/0.10

optimize their respective parameter-efficient modules; EWC and LwF apply their continual-learning regularizers; ECLIPSE and ConOVS follow their published continual-segmentation training/update protocols, adapted to the same backbone and data stream. For modular methods, completed specialists remain frozen after their increment; for non-modular methods, the model is updated sequentially on the stream. Unless otherwise specified, each increment uses the same per-increment update budget and batch size as VIRA.

Table 17: Protocol summary for the system-level comparison in Table 1. "Native" means that the method keeps its method-specific objective/update rule while sharing the common initialization, stream, no-replay constraint, and evaluation protocol.

Method family	Training/update rule	Replay	Explicit task ID at test	Deployment in this comparison
Frozen source	none	none	no	frozen base
Mask-Adapter	sequential model update	none	no	final model
LoRA-only / Adapter-only	native PEFT update	none	no	final adapted model
EWC / LwF	native regularized sequential update	none	no	final model
ECLIPSE / ConOVS	published continual update adapted to common backbone/stream	none	no	method-native continual prediction
VIRA	append specialist; freeze completed modules	none	no	top-1 retrieval + region fallback

At test time, no explicit task or dataset identifier is provided. VIRA uses its top-1 compatibility retrieval and region-wise base–specialist arbitration, whereas comparison methods use their method-native deployment rules under the same evaluation vocabularies, preprocessing, evaluator, and metrics. We report mIoU and, where available, PQ/SQ/RQ. Retention is measured on COCO, Pascal Context-59, and Pascal VOC-20; expanded-taxonomy transfer uses ADE-847 and Pascal Context-459. Training budgets, parameter counts, and storage are reported separately from accuracy.

F.3 Selector baselines

Table **??** compares vocabulary-nearest selection, global fallback, and region-level arbitration, exposing their different retention–specialization trade-offs.

F.4 Prefix-bank temporal evaluation

The main text reports the prefix-bank T1/T2/T3 trajectory and frozen-base fallback rates. Historical specialists remain parameter-wise unchanged as the bank grows; changes in earlier-domain deployed scores arise from retrieval and arbitration over the expanded candidate bank. Under forced matched-specialist activation, completed modules therefore show approximately zero backward transfer, while deployed endpoints without explicit task IDs can still move as candidate selection changes.

F.5 POST-TRAINING COEFFICIENT INTERVENTIONS

The coefficient interventions in the main text use forced matched-specialist activation and are post-training and inference-only. Specialist parameters, images, output vocabulary, and evaluator are fixed, while candidate selection and base–specialist arbitration are bypassed by activating the matched specialist directly. They therefore probe sensitivity to the learned coefficient realization rather than the contribution of dynamic composition under retraining. Fixed-mean replaces each region's coefficient vector with the dataset mean; within-image permutation reassigns coefficient vectors among regions of the same image; across-image permutation reassigns them across images; uniform sets the residual mixture to the centered uniform reference. This isolates coefficient identity and image correspondence without retraining.

F.6 MAFT ARCHITECTURE TRANSFER

The main-text MAFT table reports matched per-increment capacity together with incremental-domain and PAS-20 results. Across repeated runs, PAS-20 mIoU is 95.58 ± 0.11 for Region only and 96.04 ± 0.09 for VIRA, compared with 96.02 for the frozen source. Thus the cross-architecture improvement is obtained without increasing per-increment capacity, while this retention benchmark remains close to the frozen-source endpoint within the observed run-to-run variation.