TAG: CALIBRATING SAM3 FOR UNIVERSAL IMAGE SEGMENTATION WITHOUT DENSE SUPERVISION

Anonymous authors

Paper under double-blind review

ABSTRACT

The recent Segment Anything Model 3 (SAM3) extends powerful promptable segmentation from spatial cues to concepts, raising the prospect of a single foundation model for universal image segmentation. However, this promise often fails to translate into reliable performance on downstream segmentation tasks. In this work, we empirically identify concept recognition, rather than mask quality, as the primary bottleneck: SAM3 accurately delineates the objects it recognizes yet struggles to robustly determine which concepts are present. Driven by this, we present *Text Anchor re-Grounding* (***TAG***), a simple yet effective framework that rectifies the concept recognition of SAM3 on the target domain without dense annotations. Specifically, ***TAG*** introduces an *Anchor Residual Pathway* comprising image-conditioned, low-rank adapters to re-ground text anchors in the target distribution and modulate subsequent text-conditioned interactions. Guided solely by image-level category labels, ***TAG*** jointly optimizes binary cross-entropy with *Soft Boundary Ranking* objectives to calibrate presence confidence and distinguish present concepts from confusable negatives. Building upon this, we devise *Confidence-Weighted Distillation* to couple rich semantic context of a multimodal large language model with fine-grained spatial priors of SAM3, thereby achieving effective calibration in the absence of manual annotations. Extensive experiments demonstrate that ***TAG*** unlocks SAM3's segmentation potential with only 3.1M trainable parameters, achieving impressive performance across diverse scenarios, and notably, even surpassing leading mask-supervised methods on several metrics. *Our code will be publicly available. Anonymous github link*.

1 INTRODUCTION

Image segmentation has long stood as a fundamental pillar of visual understanding. Recent advances increasingly pursue a unified model capable of handling semantic, instance, and panoptic segmentation (Jain et al., 2023; Cheng et al., 2021b). To this end, prevailing universal frameworks (Cheng et al., 2022; Cavagnero et al., 2024) unify the three tasks under a shared mask-classification paradigm, and successive improvements in model architecture and learning objectives have steadily refined mask quality (Li et al., 2023a; 2024c). Despite remarkable progress, carrying such a model to new target domains remains challenging. On the one hand, adapting these models to a target domain relies heavily on dense pixel-level annotations, which are notoriously labor-intensive and time-consuming to obtain. On the other hand, such specialized models typically require full fine-tuning for each deployment scenario, incurring substantial computational overhead.

The Segment Anything Models (SAM) (Kirillov et al., 2023; Ravi et al., 2025) offer a promising route forward by amortizing dense mask learning through large-scale pretraining, yet their reliance on geometric prompts leaves semantic recognition external to the model. Recent SAM3 (Carion et al., 2026) marks a pivotal leap by seamlessly fusing text into its visual pipeline, extending promptable paradigm from spatial cues to concepts. In principle, a frozen SAM3 can operate as a universal image segmenter, yielding both semantic and instance masks in a single pass when guided merely by target class names. In practice, however, SAM3 falls markedly short of the promise when directly applied on downstream segmentation benchmarks. As shown in Figure 1, SAM3 exhibits two prevalent failure modes: i) *inter-class confusion*, where the model falsely activates on visually similar yet incorrect categories, and ii) *under-activation*, where clearly visible instances remain un-

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.

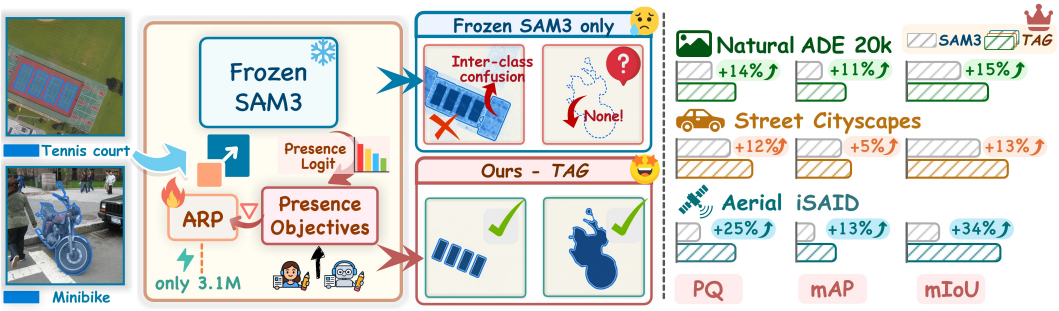

Figure 1: ***TAG* calibrates SAM3 with image-level supervision.** ***TAG*** learns an ARP with only 3.1M parameters using image-level presence objectives while keeping SAM3 fully frozen. The examples illustrate how ***TAG*** corrects the inter-class confusion and under-activation of SAM3. Right panel reports the gains achieved with image-level ground-truth labels across different scenarios.

detected. To address this, a natural way is to fine-tune SAM3 with pixel-level annotations so that it adapts to the target distribution (Carion et al., 2026). Though effective, this manner reintroduces the substantial computational burden that large-scale pretraining is intended to alleviate and requires prohibitive mask annotations, limiting its real-world deployment. This raises a central question: *How can SAM3 be calibrated at low computational and annotation cost?*

Answering the above question first requires pinpointing where SAM3 fails. Toward this end, this paper conducts an in-depth analysis of the failure modes of SAM3 (detailed results are provided in §2.2). Crucially, our empirical analysis reveals that these failures do not arise from scattered pixel-level noise, instead, they form large contiguous regions spanning entire objects, causing objects to be misclassified or missed entirely. By contrast, SAM3 accurately delineates the objects it does recognize. Even with SAM3's raw masks left unchanged, segmentation performance can be markedly improved simply by correcting its category assignments and activations, as shown in Figure 2b. *These findings converge on a clear diagnosis: the mask prior of SAM3 transfers effectively to the target domains, whereas concept-level recognition struggles to generalize.*

Motivated by this insight, we propose ***Text Anchor re-Grounding (TAG)***, a simple yet effective framework that rectifies the concept recognition of SAM3 on the target domain without mask annotations. Specifically, **on the learnable modules**, ***TAG*** introduces an *Anchor Residual Pathway* (ARP) that leverages an image-conditioned, low-rank text feature adapter to dynamically re-ground class embeddings and propagates these corrections through subsequent text-conditioned interactions, with all pretrained SAM3 weights frozen. Besides, **on the learning objectives**, ***TAG*** jointly optimizes binary cross-entropy and *Soft Boundary Ranking* solely by image-level category labels to calibrate concept presence and separate present categories from confusable negatives, correcting the semantic decisions that govern mask prediction. **To transcend reliance on manual annotations**, we devise *Confidence-Weighted Distillation* strategy that robustly harnesses category labels from multimodal large language model (MLLM), *i.e.*, Qwen3.8 (Qwen Team, 2026), to guide the learning of the ARP on unlabeled images. In this way, ***TAG*** distills the off-the-shelf semantic knowledge of MLLM into lightweight adapters, bridging two foundations with complementary strengths.

Extensive experiments on multiple benchmarks spanning natural images, urban scenes, and aerial imagery demonstrate the effectiveness and efficiency of ***TAG***. Despite introducing negligible **3.1M** trainable parameters under coarse image-level presence labels, ***TAG-I*/*TAG-V*** (human/MLLM supervised) remarkably boost the frozen SAM3 by 17.2/14.0 PQ, 9.5/7.5 AP, and 20.8/16.8 mIoU across three benchmarks. Both variants surpass all competing weakly supervised and open-vocabulary methods, *and notably*, even outperform leading mask-supervised models on key metrics.

Briefly put, our contributions are summarized as follows:

- ❶ **Bottleneck diagnosis:** Through empirical analysis and oracle interventions, we reveal that the performance bottleneck of SAM3 in downstream segmentation tasks is concept grounding rather than mask quality. Correcting semantic bias while leaving the raw masks intact can unleash the powerful dense perception capability of SAM3.
- ❷ **Targeted solution:** Guided by the diagnosis, we propose ***TAG***, which calibrates the text-conditioned process of SAM3 through an *Anchor Residual Pathway*, optimized with binary

cross-entropy and *Soft Boundary Ranking* objectives from image-level category labels. We further introduce Confidence-Weighted Distillation to robustly harness the global semantic recognition priors of MLLM, generalizing ***TAG*** to the unlabeled regime.

❸ **Broad validation:** Extensive evaluations across natural, urban scene and aerial benchmarks demonstrate the superior efficacy and efficiency of ***TAG***. With only 3.1M learnable parameters, it achieves impressive performance in both weakly supervised and annotation-free settings, even surpassing leading fully supervised methods on key metrics.

2 PRELIMINARIES

2.1 A RECAP ON SAM3

We build on the image branch of *SAM3* (omitting all video components), which segments visual concepts specified by noun phrases. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a benchmark-defined label set $\mathcal{C} = \{c_1, \ldots, c_K\}$, we use the canonical class name c_k as the text prompt. The image is first mapped to unconditioned visual tokens $\mathbf{F}_v \in \mathbb{R}^{N_v\times d}$ by image encoder, while each class name is independently encoded into text embeddings $\mathbf{F}_t^k \in \mathbb{R}^{N_t\times d}$ by text encoder.

For each class k, the fusion encoder conditions the visual tokens on the text tokens through stacked cross-attention layers. The fusion encoder is followed by a DETR-like (Carion et al., 2020) decoder that transforms a set of object queries $\mathbf{O} \in \mathbb{R}^{Q\times d}$ into class-conditioned predictions. A single learned *presence token* is concatenated with the object queries $\tilde{\mathbf{O}} = [\mathbf{O}; \mathbf{o}_{\text{pres}}] \in \mathbb{R}^{(Q+1)\times d}$ to predict whether a category appears in an image. Each decoder block applies cross-attention from $\tilde{\mathbf{O}}$ to the text embeddings and the conditioned visual tokens in turn. After decoder, the presence token produces the global presence score $p^k_{\text{pres}} = \sigma(\text{MLP}_{\text{pres}}(\mathbf{o}_{\text{pres}}))$ for each text input. Following the MaskFormer (Cheng et al., 2021b), *SAM3* produces both semantic and instance segmentation masks for each class k through a shared segmentation head. The conditioned features from the fusion encoder are used to generate the semantic segmentation mask, whereas instance segmentation additionally leverages the object queries $\mathbf{O}$ output by the decoder.

For dense prediction, we follow SegEarth-OV3 (Li et al., 2025) and adopt the presence score p^k_{pres} to weight mask predictions during filtering and fusion to suppress non-salient categories.

2.2 WHERE ZERO-SHOT SAM3 FAILS

Directly applied to downstream segmentation, SAM3 often confuses competing categories or misses visible objects. In this work, we find that errors in SAM3 predominantly manifest as holistic, object-level failures, rather than the localized pixel-level noise commonly observed in traditional segmentation models. Specifically, in Figure 2a, the mask predicted as *pillow* closely follows the contour of *cushion*, achieving **93.2%** class-agnostic IoU with its ground-truth mask. Despite this accurate delineation, the entire object is misclassified as a competing neighboring category. A complementary failure occurs in the *vehicle* example, where the confidence of the correct concept falls below the filtering threshold, causing the model to suppress the entire object as background. Additional object-level statistics and diagnostic experiments are provided in Appendix D.5. These observations motivate the following hypothesis: *SAM3 exhibits robust spatial perception but fails to recognize concepts reliably, which prevents accurate object geometry from being translated into correct dense predictions and consequently leads to suboptimal performance on downstream benchmarks.*

Oracle diagnosis of the recognition bottleneck. To investigate this hypothesis, we compare oracle interventions on region geometry, category assignment, and concept activation in the frozen model, as shown in Figure 2b. Even perfect geometry (*GT mask regions*) leaves substantial errors when regions are named by SAM3's classification scores, and reduces accuracy on ADE20K. With raw masks unchanged, *GT categories* addresses inter-class confusion and improves ADE20K by **47.2** mIoU over zero-shot SAM3. *Positive gates* counters under-activation by enabling GT-present concepts and also improves accuracy. Their combination remarkably outperforms *GT regions* on every benchmark, demonstrating the larger benefit of correcting recognition over supplying perfect geometry. These gains unlock the segmentation capacity already present in SAM3's spatial predictions but obscured by unreliable semantic recognition.

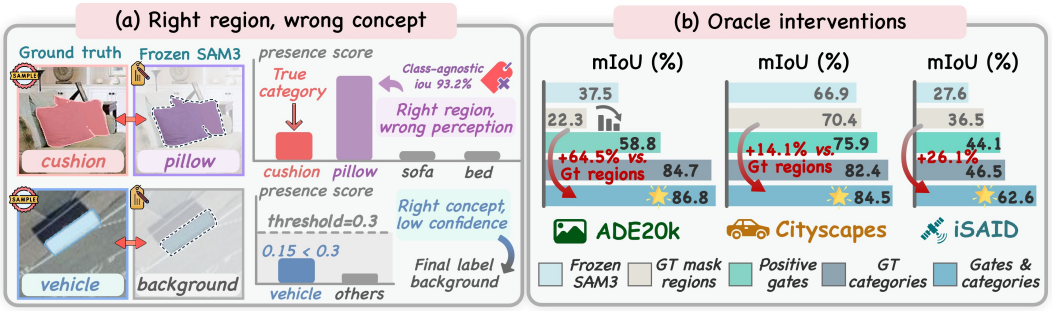

Figure 2: **Recognition errors in frozen SAM3. (a)** SAM3 exhibits strong foreground perception, yet its concept recognition often fails. **(b)** Oracle interventions: *Frozen SAM3* gives original predictions. *GT mask regions* pairs GT mask regions with SAM3's original category scores. *Positive gates* keeps raw SAM3 masks, setting only GT categories gate to 1 for instance filtering. *GT categories* pairs SAM3 masks with GT categories. *Gates & Categories* combines both semantic corrections.

Toward efficient adaptation of SAM3. Together, the object-level error pattern and the substantial gains from semantic correction support our hypothesis that SAM3 retains strong spatial priors for mask delineation, yet its category recognition fails to generalize reliably to downstream domains. It suggests an efficient way for adapting SAM3. Rather than modifying its pretrained spatial representations, we can preserve all pretrained weights to reuse the spatial priors, and concentrate the learning effort on correcting *category-level semantic recognition*. In this setting, higher-level semantic labels can provide effective supervision at relatively low annotation cost, while the pretrained model supplies the fine-grained spatial information required for dense prediction.

3 Method

We present ***TAG***, which re-grounds SAM3's semantic anchors in the target domain while keeping all pretrained weights frozen (Figure 3). We introduce the *Anchor Residual Pathway* (§3.1) and its *learning objectives* under image-level supervision (§3.2). To generalize across unlabeled images, we develop *Confidence-Weighted Distillation* (§3.3) using presence labels from an off-the-shelf MLLM.

3.1 Anchor Residual Pathway

Motivation. Within SAM3, the concept recognition is mediated by the text-conditioned process. The text features are computed once from the canonical class names and condition both cross-modal interactions and presence prediction throughout the model. However, since the text features remain static across all images, they merely inherit each category's prototypical semantics from pre-training rather than capture how it actually appears in the target domain. Under domain shift, this mismatch leads to the failures observed in Figure 1, *i.e.*, the anchor lies too close to a neighboring category (inter-class confusion) or the anchor lies too far from its own category (under-activation).

Re-grounding the anchors. Driven by this, we freeze all pretrained SAM3 weights and focus adaptation on text features and the pathways through which they condition the model. We first learn a low-rank residual (Hu et al., 2022) to adapt the text features to the target domain:

$$\Delta_{\mathrm{t}}(\mathbf{F}_t^k \mid \bar{\mathbf{e}}_k) = \big((\mathbf{F}_t^k \mathbf{A}) \odot g(\bar{\mathbf{e}}_k)\big)\mathbf{B}, \qquad g(\bar{\mathbf{e}}_k) = 1 + \tanh\big(\mathrm{MLP}(\bar{\mathbf{e}}_k)\big), \tag{1}$$

here $\bar{\mathbf{e}}_k \in \mathbb{R}^{1\times d}$ denotes the pooled text tokens, $\mathbf{A} \in \mathbb{R}^{d\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times d}$ are learnable projections shared across classes, with the rank $r = 16$. The class-specific gate $g(\bar{\mathbf{e}}_k) \in \mathbb{R}^{1\times r}$ reweights the low-rank features by element-wise multiplication.

Since the low-rank residual is image-invariant, we add *text-to-image cross-attention* to modulate text anchors with visual evidence. Specifically, we first map frozen visual features $\mathbf{F}_v$ to visual keys and values $\mathbf{U} \in \mathbb{R}^{N_v\times d}$. Their cosine similarities with the pooled text feature $\bar{\mathbf{e}}_k$, scaled by a learned factor η, form a class-specific bias $\mathbf{a}_k \in \mathbb{R}^{1\times N_v}$ added to attention logits. We refine the original text queries $\mathbf{F}_t^k$ through two trainable cross-attention blocks and form the image-conditioned residual as:

$$\Delta_{\mathrm{img}}^k(I) = \mathrm{Proj}_{\mathrm{out}}\big(\mathrm{CrossAttention}(\mathbf{F}_t^k, \mathbf{U}; \mathbf{a}_k)\big), \tag{2}$$

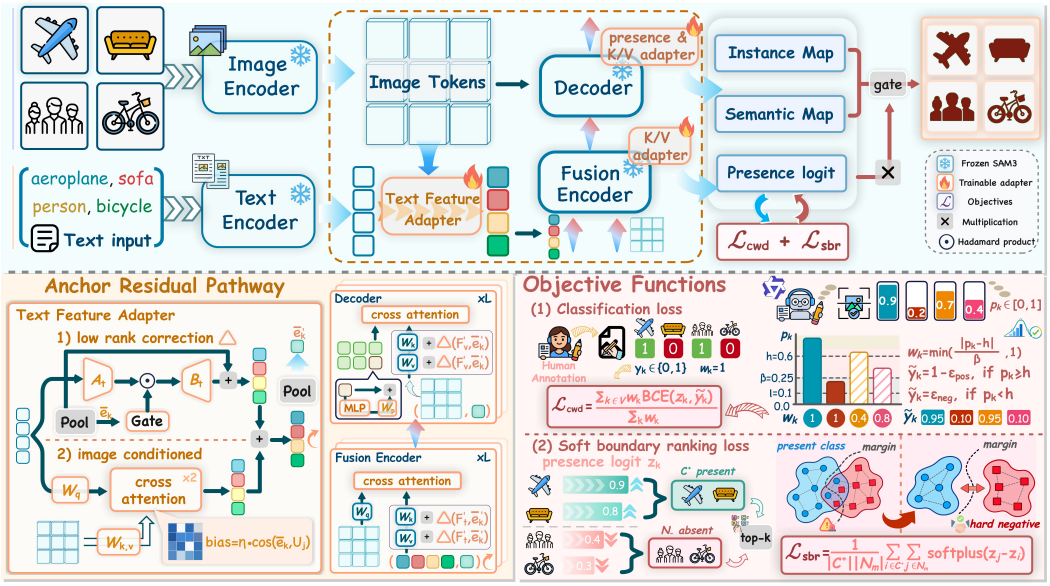

Figure 3: **Overview of our proposed *TAG*.** Frozen SAM3 encodes the image and class prompts, while trainable adapters refine text features, cross-modal attention, and presence prediction using image-level supervision. Here, ▢ and ▢ denote the object queries and presence token, respectively.

where $\mathrm{Proj}_{\mathrm{out}}$ maps the final output to an $N_t \times d$ residual. Finally, we combine the domain correction and image-conditioned residual to jointly update the original text features:

$$\mathbf{F}_t'^k(I) = \mathbf{F}_t^k + \Delta_{\mathrm{t}}(\mathbf{F}_t^k \mid \bar{\mathbf{e}}_k) + \Delta_{\mathrm{img}}^k(I). \tag{3}$$

Propagating semantic corrections. Within the fusion encoder, the corrected anchors are shared across all blocks, whereas the visual representations conditioned on them evolve progressively with network depth. To adapt this, we further introduce anchor residuals into the textual keys and values of the fusion encoder blocks. The mean corrected text feature $\bar{\mathbf{e}}'_k \in \mathbb{R}^{1\times d}$ conditions residual updates to the text keys and values $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{N_t \times d}$ before image-to-text cross-attention block:

$$\mathbf{K}' = \mathbf{K} + \Delta_K(\mathbf{F}_t'^k(I) \mid \bar{\mathbf{e}}'_k), \qquad \mathbf{V}' = \mathbf{V} + \Delta_V(\mathbf{F}_t'^k(I) \mid \bar{\mathbf{e}}'_k), \tag{4}$$

here Δ_K and Δ_V follow Eq. 1. The projection and gating parameters are shared across classes. In addition, we leverage the mean corrected text feature to provide the presence token with a class-specific initialization, facilitating its integration of category semantics:

$$\mathbf{o}_{\mathrm{pres}}'^k = \mathbf{o}_{\mathrm{pres}} + \mathrm{MLP}_{\mathrm{p}}(\bar{\mathbf{e}}'_k)\mathbf{W}_{\mathrm{p}}, \tag{5}$$

where $\mathbf{W}_{\mathrm{p}} \in \mathbb{R}^{G\times d}$ contains G learnable vectors shared across classes, and $\mathrm{MLP}_{\mathrm{p}}$ predicts their combination weights from $\bar{\mathbf{e}}'_k$. To guide object queries toward visual evidence of the intended category, each decoder layer similarly refines the image keys and values projected from the fusion output $\mathbf{F}_v^k \in \mathbb{R}^{N_v \times d}$. These residuals follow Eq. 1, using $\mathbf{F}_v^k$ as input and $\bar{\mathbf{e}}'_k$ for gating.

3.2 Learning from Image-Level Presence

Motivation. Pretrained spatial priors of SAM3 supply the cues for mask delineation, thus, the supervision required for learning can be ***coarser*** than the desired output. Inspired by this, we optimize ARP by class labels to recalibrate category recognition, avoiding labor-intensive dense annotations.

Per-class presence learning. For each image I and input class k, the decoded presence token produces a classification logit $z_k = \mathrm{MLP}_{\mathrm{pres}}(\mathbf{o}_{\mathrm{pres}})$, supervised by the binary presence label $y_k \in \{0, 1\}$. We train ARP with standard binary cross-entropy (BCE) objective on presence logits:

$$\mathcal{L}_{\mathrm{pres}} = -\frac{1}{|\mathcal{V}|}\sum_{k\in\mathcal{V}} \left[y_k \log \sigma(z_k) + (1-y_k)\log\big(1-\sigma(z_k)\big)\right], \tag{6}$$

where $\mathcal{V}$ denotes the supervised class indices in the current benchmark.

Limitation. BCE calibrates each category independently, but does not explicitly penalize pairwise ranking errors between positive classes and the most confident negatives. This distinction matters because presence scores modulate class-specific masks that subsequently compete during inference.

Soft Boundary Ranking. To explicitly separate present concepts from confusable absent alternatives, we introduce Soft Boundary Ranking (SBR), a smooth pairwise ranking objective. Let $\mathcal{C}^+ = \{k \in \mathcal{V} : y_k = 1\}$ and $\mathcal{C}^- = \{k \in \mathcal{V} : y_k = 0\}$ denote the present and absent classes, respectively. We select m absent classes with the highest logits as hard negatives $\mathcal{N}_m$ and minimize

$$\mathcal{L}_{\text{sbr}} = \frac{1}{|\mathcal{C}^+|\,|\mathcal{N}_m|} \sum_{i \in \mathcal{C}^+} \sum_{j \in \mathcal{N}_m} \text{softplus}(z_j - z_i), \tag{7}$$

where $\text{softplus}(u) = \log(1 + e^u)$ and $m = 3$ by default. Together with BCE objective, it calibrates both absolute presence confidence and relative semantic preference, refining the presence decisions that control pretrained mask filtering. Finally, the overall objective is defined as:

$$\mathcal{L}_{\text{total}} = \underbrace{\lambda_{\text{pres}}\mathcal{L}_{\text{pres}}}_{\text{Presence calibration}} + \underbrace{\lambda_{\text{rank}}\mathcal{L}_{\text{sbr}}}_{\text{Concept ranking}}. \tag{8}$$

3.3 Semantic Distillation from a Vision-Language Model

Learning ARP requires knowing which categories are present, rather than annotating their spatial extent. This image-level supervision aligns with the global semantic understanding of multimodal large language models (MLLMs), complementing SAM3's pretrained spatial priors. Motivated by this, we develop an *unlabeled* variant by replacing human-annotated presence labels with MLLM-generated pseudo labels, enabling SAM3 to be recalibrated on unlabeled target images.

Generating presence pseudo-labels. We query Qwen3.8-27B (Qwen Team, 2026) offline to identify the categories present in each training image. An example prompt is provided below:

Presence labeling prompt (summary)

This image depicts {`scene description`}. Examine the entire image and identify which categories in {`category list`} are present, considering both foreground objects and background regions.

Return only a JSON object with the listed category names as keys and `"yes"` or `"no"` as values.

The resulting pseudo-labels for each image are continuous presence scores $p_k \in [0, 1]$ for each class k, derived from the answer-token probabilities. However, optimizing the naive BCE loss on binarized MLLM outputs discards critical confidence cues by penalizing ambiguous decisions on par with confident ones. In this setting, missed positives suppress valid concepts, while hallucinated positives can reinforce spurious presence predictions.

Confidence-Weighted Distillation. To mitigate this issue, we introduce Confidence-Weighted Distillation (CWD) strategy, which attenuates the influence of noisy supervision through confidence-aware weighting. CWD downweights uncertain pseudo-labels while preserving high-confidence cues for both concept presence and absence. Given thresholds $0 \le \ell < h \le 1$, we assign a positive binary target when $p_k \ge h$ and a negative target otherwise, followed by separate smoothing:

$$\hat{y}_k = \mathbb{1}[p_k \ge h], \qquad \tilde{y}_k = (1 - \epsilon_{\text{pos}})\hat{y}_k + \epsilon_{\text{neg}}(1 - \hat{y}_k), \tag{9}$$

here $\mathbb{1}[\cdot]$ denotes the indicator function, $\hat{y}_k$ and $\tilde{y}_k$ are the binarized target and its smoothed counterpart. The parameters $\epsilon_{\text{pos}}, \epsilon_{\text{neg}} \in [0, 1/2)$ independently control the smoothing strength for positive and negative targets, respectively. We reduce the weights of pseudo-labels with presence scores close to the decision threshold h in the BCE objective function, formulated as:

$$w_k = \min\left\{1, \frac{|p_k - h|}{\beta}\right\}, \qquad \mathcal{L}_{\text{cwd}} = \frac{\sum_{k \in \mathcal{V}} w_k \,\text{BCE}(z_k, \tilde{y}_k)}{\sum_{k \in \mathcal{V}} w_k}, \tag{10}$$

where $w_k \in [0, 1]$ denotes the confidence weight for class k, and $\beta = \max\{(h - \ell)/2, 10^{-6}\}$ is the distance scale. We replace $\mathcal{L}_{\text{pres}}$ in Eq. 8 with $\mathcal{L}_{\text{cwd}}$. Meanwhile, SBR uses only the confident classes $\mathcal{V}_{\text{conf}} = \{k \in \mathcal{V} : p_k \le \ell \text{ or } p_k \ge h\}$, forming both positive set and hard-negative candidates from this set according to $\hat{y}_k$. Intuitively, ***TAG*** can be interpreted as a form of semantic distillation, transferring the semantic understanding of MLLM to mitigate SAM3's recognition bottleneck.

Table 1: **Universal segmentation comparison** on three validation benchmarks. Sup.: $\mathcal{M}$ target masks, $\widetilde{\mathcal{M}}$ source masks, $\mathcal{I}$ image labels, $\mathcal{P}$ points, $\mathcal{V}$ MLLM labels; #Par.: trainable parameters. ↑: gains over frozen SAM3. See Appendix B for more details.

Method	Sup.	Backbone	#Par.	ADE20K			Cityscapes			iSAID		
				PQ	AP	mIoU	PQ	AP	mIoU	PQ	AP	mIoU
Fully supervised universal segmentation												
Mask2Former [CVPR22]	$\mathcal{M}$	Swin-L	216M	48.1	34.9	56.1	66.6	43.7	83.3	58.1	39.5	68.5
MaskDINO [CVPR23]	$\mathcal{M}$	R50	52M	–	–	48.7	–	–	80.0	57.4	36.8	64.6
OneFormer [CVPR23]	$\mathcal{M}$	DiNAT-L	223M	50.5	36.0	58.3	67.6	45.6	83.1	51.3	31.7	62.9
PEM [CVPR24]	$\mathcal{M}$	R50	35.6M	38.5	–	45.5	61.1	–	79.9	49.1	30.7	59.7
EoMT [CVPR25]	$\mathcal{M}$	DINOv2-L	316M	50.6	–	58.4	–	–	84.2	56.7	36.3	69.7
ViT-P [CVPR26]	$\mathcal{M}$	InternImage-H	1.4B	54.5	40.6	61.2	70.8	50.6	85.4	–	–	–
Open-vocabulary universal segmentation												
X-Decoder [CVPR23]	$\widetilde{\mathcal{M}}$	DaViT-L	340.5M	21.8	13.1	29.6	38.1	24.9	52.0	28.1	17.6	37.1
HIPIE [NeurIPS23]	$\widetilde{\mathcal{M}}$	ViT-H	202.3M	22.9	19.0	29.0	–	–	–	28.6	13.2	9.5
APE [CVPR24]	$\widetilde{\mathcal{M}}$	ViT-L	380.3M	27.2	24.4	30.0	33.3	30.3	44.2	32.9	23.5	37.4
SAM-CP [ICLR25]	$\widetilde{\mathcal{M}}$	ConvNeXt-L CLIP	–	27.2	17.0	31.8	41.0	29.3	47.9	–	–	–
OpenWorldSAM [NeurIPS25]	$\widetilde{\mathcal{M}}$	SAM2 + VLM	4.5M	35.2	16.9	60.4	36.1	12.0	57.4	21.6	10.5	22.3
FrozenSeg [IJCV26]	$\widetilde{\mathcal{M}}$	ConvNeXt-L + SAM	26.5M	25.9	16.4	34.4	45.8	28.4	56.8	33.5	19.2	32.5
Weakly supervised universal segmentation												
Point2Mask [ICCV23]	$\mathcal{P}$	R50	56.5M	18.6	11.8	26.3	21.2	3.5	39.9	5.2	1.0	31.9
SAM3 [ICLR2026]	-	SAM3	*freeze*	32.7	25.0	37.5	50.3	36.5	66.9	19.2	12.2	27.6
TAG-V (Ours)	$\mathcal{V}$	SAM3	3.1M	**40.8**↑8.1	**31.5**↑6.5	**47.0**↑9.5	**61.6**↑11.3	**41.0**↑4.5	**78.1**↑11.2	**41.9**↑22.7	**23.6**↑11.4	**57.4**↑29.8
TAG-I (Ours)	$\mathcal{I}$	SAM3	3.1M	**46.9**↑14.2	**35.7**↑10.7	**52.7**↑15.2	**62.7**↑12.4	**41.0**↑4.5	**80.3**↑13.4	**44.2**↑25.0	**25.5**↑13.3	**61.4**↑33.8

4 EXPERIMENTS

Benchmarks. We conduct our main experiments on three benchmarks spanning natural images **ADE20K** (Zhou et al., 2017), urban street scenes **Cityscapes** (Cordts et al., 2016), and aerial imagery **iSAID** (Waqas Zamir et al., 2019). Compared with Pascal VOC (Everingham et al., 2015) and COCO (Caesar et al., 2018), two widely used segmentation benchmarks, we prioritize the more challenging datasets described above to enable a rigorous evaluation of cross-domain adaptation. More comparative results on Pascal VOC and COCO are provided in Appendix D.2. Following standard universal image segmentation evaluation (Cheng et al., 2022), we report panoptic quality (PQ) (Kirillov et al., 2019), mask average precision (AP), and mean intersection-over-union (mIoU) for panoptic, instance, and semantic segmentation tasks, respectively.

Implementation details. Our work builds upon the pretrained SAM3 (Carion et al., 2026), which employs Perception Encoder (Bolya et al., 2025) for image and text encoding. As for text input, we directly use the canonical class names defined by each benchmark as concept prompts. We adopt the open-source Qwen3.8 27B (Qwen Team, 2026) as the default MLLM for generating labels in unlabeled setting. Our implementation is based on `Pytorch`, and all experiments are conducted on NVIDIA RTX 4090 GPUs. More implementation details are provided in Appendix C.

4.1 COMPARATIVE RESULTS

Counterparts. Achieving universal image segmentation under weak supervision represents a formidable yet nascent frontier. While pioneering efforts (Shen et al., 2021; Li et al., 2023b) have made commendable strides, their efficacy remains limited on challenging scenarios. In light of this, we expand the comparison to leading fully supervised and open-vocabulary methods, positioning ***TAG*** within the broader landscape of segmentation research (see Appendix B for more details).

Strong priors make weak supervision effective. Table 1 compares ***TAG*** with weakly supervised methods on challenging cross-domain datasets. Despite utilizing point annotations, these methods still remain suboptimal, whereas ***TAG*** achieves impressive results relying solely on image-level presence. ***TAG*** inherits the spatial priors of SAM3 and directs weak supervision solely toward category recognition, providing a viable way for efficient and effective universal image segmentation.

Optimal trade-off for application. While open-vocabulary models can generalize to novel categories, they fail under severe domain shifts (Appendix B.2). These models rely on mask-supervised adaptation within the related domain, which substantially limits the deployment advantages offered by vocabulary flexibility. With only MLLM-generated labels, ***TAG-V*** surpasses all open-vocabulary

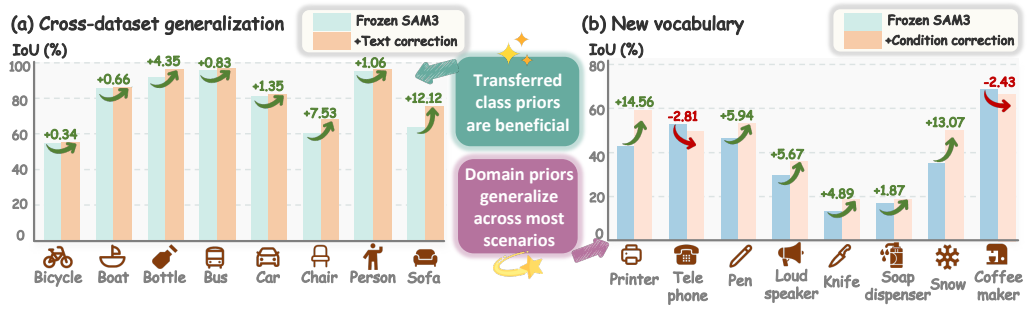

Figure 4: **Transferability of *TAG*.** (a) Text residuals trained on ADE20K and generalized to shared VOC21 classes. (b) Class-agnostic adapters trained on ADE-150 and transferred to new classes.

Table 2: **Learnable module ablations.** PQ (%); w/o: without.

Setting	ADE	City	iSAID	**Avg.**	Δ
w/o feature adapter	43.88	59.81	42.11	48.60	−2.66
w/o fusion K/V	44.52	60.47	42.34	49.11	−2.15
w/o decoder K/V	45.18	61.38	42.62	49.73	−1.53
w/o presence token	45.99	61.36	43.74	50.36	−0.89
All adapters	**46.86**	**62.67**	**44.24**	**51.26**	–

Table 3: **Objective function ablations.** PQ (%) for ***TAG-I*** and ***TAG-V***.

Setting	ADE	City	iSAID	**Avg.**	Δ
w/o SBR	44.82	60.25	41.85	48.97	−2.28
w/o BCE	43.48	57.77	41.66	47.64	−3.62
TAG-I	**46.86**	**62.67**	**44.24**	**51.26**	–
naive BCE	37.54	57.65	38.59	44.59	−3.52
TAG-V	**40.80**	**61.59**	**41.95**	**48.11**	–

counterparts across nine metrics, except the mIoU of OpenWorldSAM on ADE20K. For tasks with predefined vocabularies, these results favor ***TAG*** as a more robust and practical alternative.

Lightweight adaptation narrows the performance gap. With only 3.1M trainable parameters, our method substantially narrows the performance gap between SAM3 and leading fully supervised methods. Notably, ***TAG-I*** even surpasses several counterparts, achieving 35.7 AP on ADE20K ahead of Mask2Former (*vs.*34.9), while attaining 61.4 mIoU on iSAID over PEM (*vs.*59.7). These results demonstrate that our method is highly competitive and establishes a strong baseline. Future advances hold the potential to match or even exceed fully supervised SOTA under weak supervision.

4.2 Diagnostic Experiments

Key Trainable Component Analysis. Table 2 shows that all four components contribute, as removing any one lowers PQ across the three datasets. The largest degradation without the feature adapter highlights the central role of semantic anchor correction. K/V and presence-token adapters further improve how these anchors guide cross-modal interaction and concept activation. Their combined effectiveness supports ARP as a coordinated pathway for semantic re-grounding.

Objective Functions. Table 3 supports the complementary roles of BCE and SBR. SBR alone still improves over frozen SAM3, suggesting that relative concept ranking can partially alleviate its recognition bias. Combining both objectives yields the best results, coupling absolute presence confidence with relative concept ordering. For ***TAG-V***, CWD consistently outperforms naive BCE, highlighting the importance of accounting for label uncertainty when transferring MLLM semantics.

Generalization Analysis. To further investigate the generalization of ***TAG***'s learned semantic priors, we conduct two transfer experiments. Specifically, we *i)* directly transfer the learned class-specific corrections to related domains, and *ii)* reuse the class-agnostic parameters with an expanded vocabulary. The quantitative results are reported in Figure 4. **As for cross-dataset generalization,** ADE20K-trained text residuals improve all shared VOC21 classes, with gains ranging from modest to pronounced. **As for vocabulary expansion,** reusing the image-conditioned, fusion, and decoder adapters trained on ADE150 yields substantial gains for most of the eight newly added ADE847 categories. These observations reveal that *i)* the category corrections can capture recognition biases shared across related domains, and *ii)* class-agnostic adapters learn visual priors shared across categories within the domain. However, given that these gains are inconsistent and occasionally negative, we argue that efficient adaptation to the target domain remains the most reliable choice.



Figure 5: **Qualitative comparison** results across natural images, urban scenes and aerial imagery.

Targeted adaptation matters beyond low-rank capacity. To explore whether generic low-rank adaptation explains the gains of our work, we compare ***TAG-I*** against ordinary LoRA under identical supervision and objectives (see Appendix C.2 for more details), as shown in Figure 6. Although generic low-rank adaptation is already effective, text feature correction alone outperforms LoRA with a parameter budget matched to the full TAG by 4.2% PQ on ADE20K, while completing the full ARP yields a further 4.1% gain. Besides, adding low-rank updates to the image features on top of TAG leads to a 3.8% performance drop. These results suggest that, under the current paradigm, the number of trainable parameters is not the key factor. What matters more is where adaptation is applied and how it is conditioned. Guided by our empirical diagnostics and architectural insights into SAM3, ARP deliberately confines its intervention to category recognition and delivers the most substantial performance gains across all adaptation strategies.

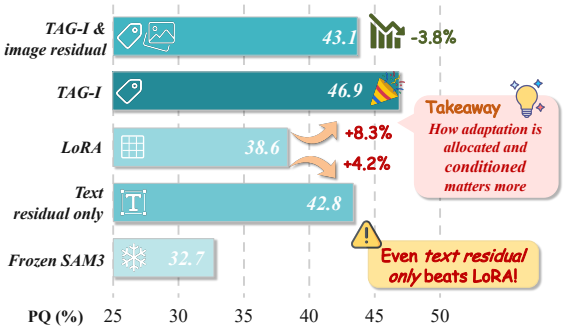

Figure 6: **Adaptation strategies on ADE20K.** PQ comparison of ***TAG-I***, LoRA, and residual variants.

4.3 Qualitative Analysis

Figure 5 provides qualitative comparisons on three benchmarks. Frozen SAM3 shows pronounced inter-class confusion and spurious activations, mislabeling objects and extending foreground predictions into background regions. Through semantic anchor correction and presence calibration, ***TAG*** alleviates these errors while retaining SAM3's spatial detail. Both ***TAG-V*** and ***TAG-I*** yield more semantically consistent masks with cleaner object extents and less background leakage. Further comparisons across diverse scene layouts and object scales appear in Appendix E.

5 Conclusion

In this paper, we present a systematic investigation into SAM3's performance breakdowns on universal image segmentation, pinpointing concept recognition as the governing bottleneck for downstream adaptation. Motivated by this insight, we introduce ***TAG*** to calibrate the text-conditioning process of SAM3, thereby unlocking the latent potential of its powerful segmentation capabilities. We further design *CWD* mechanism to effectively bridge high-level semantic context from MLLM with the precise spatial representations of SAM3, enabling semantic calibration in the unlabeled setting. Empirical results demonstrate the efficiency and effectiveness of ***TAG***, which achieves impressive segmentation performance with only 3.1M trainable parameters. Our work establishes a strong baseline for challenging universal image segmentation under weak supervision and label-free regimes, paving the way to match or exceed fully supervised SOTAs.

AI USE STATEMENT

We used a locally deployed, open-source Qwen3.8 27B to generate image-level category presence labels. Besides, we used GPT to improve the language and readability of the manuscript. All AI-assisted outputs were carefully reviewed and verified by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide a detailed description of ***TAG*** and its learning objectives in Section 3. The Implementation Details in Section 4 and Appendix C.1 specify the experimental setup, optimization procedure, and training hyperparameters. We also provide our model code through an anonymous link accompanying the submission.

REFERENCES

Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019.

Abhishek Aich, Yumin Suh, Samuel Schulter, and Manmohan Chandraker. Progressive token length scaling in transformer encoders for efficient universal segmentation. In *ICLR*, 2025.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. In *NeurIPS*, 2025.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *ICLR*, 2026.

Niccolò Cavagnero, Gabriele Rosi, Claudia Cuttano, Francesca Pistilli, Marco Ciccone, Giuseppe Averta, and Fabio Cermelli. PEM: Prototype-based efficient MaskFormer for image segmentation. In *CVPR*, 2024.

Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 2024a.

Pengfei Chen, Lingxi Xie, Xinyue Huo, Xuehui Yu, Xiaopeng Zhang, Yingfei Sun, Zhenjun Han, and Qi Tian. SAM-CP: Marrying SAM with composable prompts for versatile segmentation. In *ICLR*, 2025.

Xi Chen, Haosen Yang, Sheng Jin, Xiatian Zhu, and Hongxun Yao. FrozenSeg: Harmonizing frozen foundation models for open-vocabulary segmentation. *arXiv preprint arXiv:2409.03525*, 2024b.

Yanhui Chen, Baoyao Yang, Siqi Liu, and Jingchao Wang. CoCo-SAM3: Harnessing concept conflict in open-vocabulary semantic segmentation. *arXiv preprint arXiv:2604.19648*, 2026.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary IoU: Improving object-centric image segmentation evaluation. In *CVPR*, 2021a.

Bowen Cheng, Alexander Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021b.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pp. 1290–1299, 2022.

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*, 2016.

Daan de Geus and Gijs Dubbelman. Task-aligned part-aware panoptic segmentation through joint object-part representations. In *CVPR*, 2024.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 2015.

Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Pointly-supervised panoptic segmentation. In *ECCV*, 2022.

Xiuye Gu, Yin Cui, Jonathan Huang, Abdullah Rashwan, Xuan Yang, Xingyi Zhou, Golnaz Ghiasi, Weicheng Kuo, Huizhong Chen, Liang-Chieh Chen, and David A Ross. DaTaSeg: Taming a universal multi-dataset multi-task segmentation model. In *NeurIPS*, 2023.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

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 *ICLR*, 2022.

Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One transformer to rule universal image segmentation. In *CVPR*, 2023.

Sanghyun Jo, Ziseok Lee, Wooyeol Lee, Jonghyun Choi, Jaesik Park, and Kyungsu Kim. TRACE: Your diffusion model is secretly an instance edge detector. In *ICLR*, 2026.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023.

Tommie Kerssies, Niccolò Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your ViT is secretly an image segmentation model. In *CVPR*, 2025.

Beomyoung Kim, YoungJoon Yoo, Chaeeun Rhee, and Junmo Kim. Beyond semantic to instance segmentation: Weakly-supervised instance segmentation via semantic knowledge transfer and self-refinement. In *CVPR*, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019.

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 *ICCV*, 2023.

Hyeokjun Kweon and Kuk-Jin Yoon. From SAM to CAMs: Exploring segment anything model for weakly supervised semantic segmentation. In *CVPR*, 2024.

Hyeokjun Kweon and Kuk-Jin Yoon. WISH: Weakly supervised instance segmentation using heterogeneous labels. In *CVPR*, 2025.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask DINO: Towards a unified transformer-based framework for object detection and segmentation. In *CVPR*, 2023a.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. Segment and recognize anything at any granularity. In *ECCV*, 2024a.

Jing Li, Junsong Fan, and Zhaoxiang Zhang. Point-supervised panoptic segmentation via estimating pseudo labels from learnable distance. In *ECCV*, 2024b.

Kaiyu Li, Shengqi Zhang, Yujie Wang, Yupeng Deng, Zhi Wang, Deyu Meng, and Xiangyong Cao. SegEarth-OV3: Exploring SAM 3 for open-vocabulary semantic segmentation in remote sensing images. *arXiv preprint arXiv:2512.08730*, 2025.

Qizhu Li, Anurag Arnab, and Philip HS Torr. Weakly- and semi-supervised panoptic segmentation. In *ECCV*, 2018.

Wentong Li, Yuqian Yuan, Song Wang, Jianke Zhu, Jianshu Li, Jian Liu, and Lei Zhang. Point2Mask: Point-supervised panoptic segmentation via optimal transport. In *ICCV*, 2023b.

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 *CVPR*, 2024c.

Zecheng Li, Zening Zeng, Yuqi Liang, and Jin-Gang Yu. Complete instances mining for weakly supervised instance segmentation. In *IJCAI*, 2023c.

Yang Liu, Yufei Yin, Chenchen Jing, Muzhi Zhu, Hao Chen, Yuling Xi, Bo Feng, Hao Wang, Shiyu Li, and Chunhua Shen. Unified open-world segmentation with multi-modal prompts. In *ICCV*, 2025.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

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*, 2014.

Dantong Niu, Xudong Wang, Xinyang Han, Long Lian, Roei Herzig, and Trevor Darrell. Unsupervised universal image segmentation. In *CVPR*, 2024.

Gensheng Pei, Xiruo Jiang, Yazhou Yao, Xiangbo Shu, Fumin Shen, and Byeungwoo Jeon. Taming SAM3 in the wild: A concept bank for open-vocabulary segmentation. *arXiv preprint arXiv:2602.06333*, 2026.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.

Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, and Xingang Wang. FreeSeg: Unified, universal and open-vocabulary image segmentation. In *CVPR*, 2023.

Qwen Team. Qwen3.8-27B. https://huggingface.co/Qwen/Qwen3.8-27B, 2026. Model card.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. In *ICLR*, 2025.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *CVPR*, 2023.

Sajjad Shahabodini, Mobina Mansoori, Farnoush Bayatmakou, Jamshid Abouei, Konstantinos N Plataniotis, and Arash Mohammadi. The missing point in vision transformers for universal image segmentation. In *CVPR*, 2026.

Yunhang Shen, Liujuan Cao, Zhiwei Chen, Feihong Lian, Baochang Zhang, Chi Su, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Toward joint thing-and-stuff mining for weakly supervised panoptic segmentation. In *CVPR*, 2021.

Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. Aligning and prompting everything all at once for universal visual perception. In *CVPR*, 2024.

Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications (SPIE)*, 2019.

Yanghui Song, Nanqing Liu, Haonan Yin, Yingjie Gao, Chengfu Yang, and Qi Ming. Prompt-calibrated SAM 3 for open-vocabulary remote sensing semantic segmentation. *IEEE Geoscience and Remote Sensing Letters*, 23, 2026.

Tran Dinh Tien and Zhiqiang Shen. ActiveSAM: Image-conditional class pruning for fast and accurate open-vocabulary segmentation. *arXiv preprint arXiv:2606.16996*, 2026.

Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model. In *NeurIPS*, 2023a.

Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. In *NeurIPS*, 2023b.

XuDong Wang, Jingfeng Yang, and Trevor Darrell. Segment anything without supervision. In *NeurIPS*, 2024.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *CVPR Workshops*, 2019.

Shiting Xiao, Rishabh Kabra, Yuhang Li, Donghyun Lee, Joao Carreira, and Priyadarshini Panda. OpenWorldSAM: Extending SAM2 for universal image segmentation with language prompts. In *NeurIPS*, 2025.

Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. EfficientSAM: Leveraged masked image pretraining for efficient segment anything. In *CVPR*, 2024.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023.

Shilin Xu, Haobo Yuan, Qingyu Shi, Lu Qi, Jingbo Wang, Yibo Yang, Yining Li, Kai Chen, Yunhai Tong, Bernard Ghanem, Xiangtai Li, and Ming-Hsuan Yang. RMP-SAM: Towards real-time multi-purpose segment anything. In *ICLR*, 2025.

Xiaobo Yang and Xiaojin Gong. Foundation model assisted weakly supervised semantic segmentation. In *WACV*, 2024.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional CLIP. In *NeurIPS*, 2023.

Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. Open-vocabulary SAM: Segment and recognize twenty-thousand classes interactively. In *ECCV*, 2024.

Bingfeng Zhang, Siyue Yu, Yunchao Wei, Yao Zhao, and Jimin Xiao. Frozen CLIP: A strong backbone for weakly supervised semantic segmentation. In *CVPR*, 2024.

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.

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 *ICCV*, 2023b.

Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-Net: Towards unified image segmentation. In *NeurIPS*, 2021.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

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

Hao Zhu, Shuo Jin, Wenbin Liao, Jiayu Xiao, Yan Zhu, Siyue Yu, and Feng Dai. VIP: Visual-guided prompt evolution for efficient dense vision-language inference. *arXiv preprint arXiv:2605.12325*, 2026.

Lianghui Zhu, Junwei Zhou, Yan Liu, Xin Hao, Wenyu Liu, and Xinggang Wang. WeakSAM: Segment anything meets weakly-supervised instance-level recognition. In *ACM MM*, 2024.

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 *CVPR*, 2023a.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *NeurIPS*, 2023b.

The appendix is divided into several sections, each giving extra information and details.

A Related Work **15**

B Counterparts **16**
B.1 Fully Supervised Universal Segmentation 17
B.2 Open-Vocabulary Universal Segmentors 17
B.3 Weakly Supervised Methods 18

C Detailed Experimental Setup **18**
C.1 Implementation Details 18
C.2 Targeted Adaptation Architectures 19

D More Quantitative Experiments **20**
D.1 Ablation Studies 20
D.2 Experiments on VOC and COCO 24
D.3 Comparison with SAM3 Calibration Methods 26
D.4 Learned Conditioning versus Retrieved Class Descriptions 26
D.5 Diagnostic Experiments 27
D.6 New-Vocabulary Evaluation on ADE847 29

E Additional Qualitative Results **30**

F Dataset-Specific Prompts for Presence Labeling **33**

A. Related Work

Universal image segmentation seeks a common framework for semantic (Long et al., 2015), instance (He et al., 2017), and panoptic (Kirillov et al., 2019) segmentation. Building on DETR's set prediction (Carion et al., 2020), MaskFormer (Cheng et al., 2021b) established mask classification as a unifying formulation, subsequently advanced through richer query–mask interaction (Cheng et al., 2022; Zhang et al., 2021) and joint cross-task training (Jain et al., 2023). MaskDINO (Li et al., 2023a) shares query representations between detection and segmentation, while OMG-Seg (Li et al., 2024c) extends task unification across image, video, open-vocabulary, and interactive segmentation. TAPPS (de Geus & Dubbelman, 2024) extends the structural granularity of panoptic segmentation by jointly representing objects and their parts with shared queries, explicitly linking each part prediction to its parent object. Efficiency is another focus: PRO-SCALE (Aich et al., 2025) reduces encoder computation by progressively introducing finer-scale feature tokens at deeper layers, while RMP-SAM (Xu et al., 2025) uses dynamic convolution and decoupled prompt/object adapters for real-time interactive, panoptic, and video instance segmentation. Other advances use a plain vision transformer (Kerssies et al., 2025) or decouple mask classification from mask generation (Shahabodini et al., 2026). These advances broaden task coverage and improve efficiency under dense supervision. Building on this progress, ***TAG*** dispenses with dense annotations for downstream adaptation and uses only low-cost image-level labels to train lightweight adapters while keeping SAM3's pretrained weights frozen. Reusing SAM3's pretrained spatial priors across all three tasks yields a favorable trade-off between annotation cost and segmentation performance.

Open-vocabulary universal segmentation extends task unification to categories specified by text at inference. X-Decoder (Zou et al., 2023a) integrates language with mask prediction, while OpenSeeD (Zhang et al., 2023b) jointly learns from segmentation and detection data through a shared semantic space and decoding mechanisms that reconcile their annotation differences. ODISE (Xu et al., 2023) transfers representations from text-to-image diffusion, and FC-CLIP (Yu et al., 2023) shares a frozen convolutional CLIP backbone between mask generation and recognition, avoiding separate image encoders for the two operations. FreeSeg (Qin et al., 2023) combines

shared segmentation parameters with adaptive task and category prompts to support all three tasks in one model. COSINE (Liu et al., 2025) unifies open-vocabulary and in-context segmentation through a shared decoder that aligns image features with textual and visual prompts. HIPIE (Wang et al., 2023b) further models whole objects, parts, and subparts hierarchically, with decoupled text–image fusion for thing and stuff categories. These approaches establish flexible category interfaces using source-domain mask supervision. Despite their ability to generalize to novel categories, direct transfer can fail under severe visual domain shift, as shown by our natural-to-aerial transfer results on iSAID (Appendix B.2, Table 4). Fine-tuning with masks from SIOR, an aerial domain closely related to iSAID, improves transfer, but obtaining dense annotations even in a related domain remains a barrier to real-world deployment. ***TAG*** requires no dense annotations for adaptation and achieves strong performance even on unlabeled target images by learning from MLLM-generated image-level labels (Table 1). This makes ***TAG*** a more practical choice for closed-set applications with a predefined target vocabulary and limited annotation resources.

Label-efficient universal segmentation reduces the need for dense target annotations. Progress in weakly supervised semantic (Zhou et al., 2016; Ru et al., 2023) and instance (Ahn et al., 2019) segmentation has extended to broader task coverage with bounding boxes (Li et al., 2018; Gu et al., 2023), points in PSPS (Fan et al., 2022) and Point2Mask (Li et al., 2023b), and image-level labels in JTSM (Shen et al., 2021). These methods demonstrate the value of sparse supervision, while the limited spatial priors available during training can constrain the recovery of detailed boundaries and instance structure. TRACE (Jo et al., 2026) makes substantial progress by extracting instance boundaries from pretrained diffusion models and combining them with semantic predictions for tag-supervised panoptic segmentation. Unsupervised approaches also advance mask discovery through self-supervised feature clustering (Niu et al., 2024) and divide-and-conquer grouping (Wang et al., 2024). ***TAG*** contributes to this direction by reusing SAM3's segmentation priors and learning semantic calibration from image-level labels. Appendix Table 10 compares these settings on VOC and COCO, including the published DHR+TRACE results.

Segment Anything Models have shifted the field toward large-scale promptable pretraining. Built on the billion-scale SA-1B corpus, SAM (Kirillov et al., 2023) translates geometric prompts into high-quality class-agnostic masks, and SAM2 (Ravi et al., 2025) carries the paradigm to video through streaming memory. Subsequent variants sharpen boundaries (Ke et al., 2023), accelerate inference (Zhang et al., 2023a; Xiong et al., 2024), or specialize to remote sensing (Chen et al., 2024a). The class-agnostic formulation of SAM and SAM2 makes them useful spatial foundations for subsequent semantic integration. Follow-up efforts couple them with grounded detectors (Ren et al., 2024), multi-modal decoders (Zou et al., 2023b), granularity-aware prompts (Li et al., 2024a), CLIP-based recognition (Yuan et al., 2024), or language prompts encoded by an external vision-language model (Xiao et al., 2025). SAM3 (Carion et al., 2026) incorporates textual supervision during pretraining and uses a dedicated presence token to support concept recognition alongside localization. ***TAG*** builds on this text-conditioned pathway for target-domain calibration.

Adapting SAM3 to target domains is an emerging pursuit, as naive application often delivers suboptimal performance. Mask-supervised fine-tuning Carion et al. (2026) specializes SAM3 through pixel-level target annotations. Among training-free methods, SegEarth-OV3 Li et al. (2025) and ActiveSAM Tien & Shen (2026) harness presence estimates to filter absent categories, while CoCo-SAM3 Chen et al. (2026) and ProC-SAM3 Song et al. (2026) enrich category prompts with linguistic variants and reconcile the ensuing predictions. Straddling both routes, ConceptBank Pei et al. (2026) distills target-calibrated class names from mask-annotated visual supports and LLM-expanded prompts. These approaches improve concept prompting and prediction fusion. ***TAG*** learns corrections throughout the text-conditioned recognition pathway using image-level presence supervision. Our diagnosis identifies substantial gains recoverable through category and activation correction with raw spatial predictions fixed, motivating this adaptation strategy. Lightweight adapters then align text-conditioned interactions with the target domain while preserving all pretrained weights. We directly compare ***TAG*** with ConceptBank, CoCo-SAM3, ActiveSAM, and ProC-SAM3 in Appendix D.3 (Table 11).

B. Counterparts

This section describes the counterparts in Table 1, their supervision settings, and the sources of the reported results. We distinguish training with target-domain masks from transfer using source-domain masks, and specify how each setting is evaluated on aerial imagery.

B.1. Fully Supervised Universal Segmentation

We compare with Mask2Former (Cheng et al., 2022), MaskDINO (Li et al., 2023a), OneFormer (Jain et al., 2023), PEM (Cavagnero et al., 2024), EoMT (Kerssies et al., 2025), and ViT-P (Shahabodini et al., 2026). These methods advance universal segmentation through query-based mask prediction, joint task modeling, efficient decoding, and pretrained vision transformers. In the fully supervised setting, a model is trained with ground-truth masks from the target domain and evaluated on held-out images from that same domain. These results therefore provide references with stronger supervision than ***TAG***.

Among our three evaluation benchmarks, the original papers primarily report results on ADE20K and Cityscapes. For these two datasets, Table 1 quotes the published scores, retaining the reported task-specific configurations and leaving unreported metrics blank. To extend the comparison to aerial imagery, we reproduce Mask2Former, MaskDINO, OneFormer, PEM, and EoMT using their official repositories, train them with iSAID training masks, and evaluate them on the iSAID validation set. ViT-P is included as a published reference on ADE20K and Cityscapes; no iSAID reproduction is reported.

B.2. Open-Vocabulary Universal Segmentors

We compare with X-Decoder (Zou et al., 2023a), HIPIE (Wang et al., 2023b), APE (Shen et al., 2024), SAM-CP (Chen et al., 2025), OpenWorldSAM (Xiao et al., 2025), and FrozenSeg (Chen et al., 2024b). These methods learn segmentation from source-domain masks and use language-conditioned recognition to predict categories beyond the source vocabulary, without requiring target-domain masks. A common protocol trains on COCO panoptic annotations and evaluates on other natural-image benchmarks, including ADE20K and Cityscapes; some methods additionally use detection, grounding, or image–text data. We quote the published natural-image results in Table 1 and evaluate the released pretrained models on aerial imagery.

Direct transfer under domain shift. Transfer from natural images to aerial imagery introduces a much larger change in viewpoint, object scale, and scene appearance. Table 4 evaluates direct transfer to iSAID using the full target vocabulary. Across the five evaluated methods, direct-transfer scores range from 1.32 to 6.05 PQ, 0.04 to 3.09 mask AP, and 3.56 to 8.08 mIoU. These measurements characterize the challenge of transferring natural-image representations to overhead scenes and motivate examining adaptation with related-domain imagery.

Table 4: **Open-vocabulary transfer to iSAID before and after SIOR adaptation.** Direct transfer uses models pretrained on natural-image data; adaptation uses SIOR masks, without iSAID training masks. Both settings evaluate the full iSAID vocabulary on 11,644 validation tiles; per-image GT class lists are never supplied. The adapted scores correspond to the iSAID open-vocabulary entries in Table 1. AP denotes mask AP. Bold marks the best score within each setting.

		Direct transfer			After SIOR adaptation		
Method	**Backbone**	**PQ**	**AP**	**mIoU**	**PQ**	**AP**	**mIoU**
X-Decoder (Zou et al., 2023a)	DaViT-L	1.32	0.33	4.78	28.09	17.61	37.10
HIPIE (Wang et al., 2023b)	ViT-H	2.28	0.57	5.53	28.56	13.17	9.53
APE (Shen et al., 2024)	EVA-02 ViT-L	**6.05**	**3.09**	7.62	32.86	**23.54**	**37.36**
OpenWorldSAM (Xiao et al., 2025)	SAM2-L + BEiT-3	2.22	0.04	3.56	21.6	10.5	22.3
FrozenSeg (Chen et al., 2024b)	ConvNeXt-L + SAM	3.26	2.97	**8.08**	**33.54**	19.18	32.53

SIOR-to-iSAID evaluation. To provide a source domain with more closely related visual characteristics, we adapt the methods on SIOR (Wang et al., 2023a), a separate aerial dataset, using its dense mask annotations and each method's official implementation. We then evaluate on iSAID with the full target vocabulary and no iSAID training masks. This preserves the source-to-target transfer protocol while providing mask supervision from a domain closer to the test images. The improvements in Table 4 demonstrate the benefit of this related-domain supervision; for example, OpenWorldSAM improves from 2.22 to 21.6 PQ after SIOR adaptation. Accordingly, the iSAID open-vocabulary entries in Table 1 report these SIOR-adapted results, and their trainable parameter counts refer to SIOR fine-tuning. The $\widetilde{\mathcal{M}}$ notation distinguishes this source-mask supervision from ***TAG***'s target image-level supervision.

B.3. Weakly Supervised Methods

JTSM (Shen et al., 2021), PSPS (Fan et al., 2022), and Point2Mask (Li et al., 2023b) establish important steps toward universal segmentation with image-level or point annotations, primarily on VOC and COCO. Point2Mask uses semantic and boundary cues within an optimal-transport formulation to derive dense supervision from sparse points. Although these approaches have advanced weakly supervised segmentation, learning detailed spatial structure remains challenging with limited spatial priors and no pretrained foundation-model mask decoder. Our Point2Mask reproductions with an R50 backbone cover ADE20K, Cityscapes, and iSAID, extending the evaluation across natural images, urban scenes, and aerial imagery (Table 1). On ADE20K, Point2Mask achieves 18.6 PQ, 11.8 mask AP, and 26.3 mIoU under point supervision. The comparison motivates studying how stronger pretrained spatial priors can complement weak semantic supervision. TRACE (Jo et al., 2026) makes substantial progress in this direction by extracting instance boundaries from diffusion-model representations. Combined with DHR semantics, it achieves strong tag-supervised panoptic results on VOC and COCO. Table 10 provides a direct comparison with the published DHR+TRACE results on VOC and COCO, alongside other weakly supervised methods including EPLD (Li et al., 2024b). This comparison uses the stated panoptic benchmarks and retains each method's reported supervision and backbone. Appendix D.2 details the protocols and the gains achieved by ***TAG***.

C. Detailed Experimental Setup

This section specifies the implementation and hyperparameters of ***TAG*** and the architectures used in the targeted-adaptation comparison.

C.1. Implementation Details

Shared training configuration. We use the same training hyperparameters on all datasets without dataset-specific tuning, as summarized in Table 5. All models are trained for 10 epochs with a peak learning rate of 5×10^{-3}, presence weight $\lambda_{\text{pres}} = 4$, ranking weight $\lambda_{\text{rank}} = 0.5$, hard-negative count $m = 3$, adapter rank $r = 16$, and $G = 16$ shared presence directions. With this common configuration, ***TAG*** consistently improves on frozen SAM3 across the evaluated domains in Table 1. These results support the generality of the adaptation strategy across natural, urban, and aerial imagery without per-dataset hyperparameter optimization. Appendix D.1.3 further examines sensitivity to objective weights, negative mining, and adapter capacity.

Table 5: **Shared training hyperparameters across datasets.** The same configuration is used without dataset-specific tuning. The loss weights follow Eq. 8, with CWD replacing BCE for ***TAG-V***. Epochs denote the scheduled training budget.

Setting	Shared value
Learning rate	5×10^{-3}
Training epochs	10
Presence weight λ_{pres}	4
Ranking weight λ_{rank}	0.5
Hard negatives m	3
Adapter rank r	16
Presence directions G	16

Repeated runs and stability. For each dataset, the ***TAG-I*** and ***TAG-V*** scores in Table 1 are arithmetic means over three independent training runs with different random seeds and otherwise identical settings. We observe little variation across runs, with a maximum sample variance of 0.3 squared percentage points across the reported dataset–metric combinations. These small fluctuations indicate stable performance across the evaluated seeds.

Model and optimization. We implement ***TAG*** in PyTorch with the pretrained image branch of SAM3 (Carion et al., 2026) and conduct experiments on NVIDIA RTX 4090 GPUs. The image and text encoders, fusion encoder, transformer decoder, and prediction heads retain their pretrained weights; only ARP parameters are optimized. We use Adam (Kingma & Ba, 2015) with weight decay 10^{-4}, numerical stabilizer 10^{-8}, and gradient-norm clipping at 0.1. The cosine OneCycle schedule (Smith & Topin, 2019) warms up for one epoch, starting at $1/25$ of the learning rate in Table 5 and ending at $1/10^4$ of that initial rate. Its first-moment coefficient cycles between 0.85

and 0.95, while the second-moment coefficient is fixed at 0.999. The learning rates of the presence, text-feature, fusion K/V, and decoder K/V parameter groups are scaled by 0.1, 0.1, 1.0, and 0.5, respectively. The forward computation uses bfloat16 autocasting, with loss evaluation in float32.

Loss settings. For ***TAG-I***, the training objective combines BCE and SBR using the shared weights in Table 5. ***TAG-V*** replaces BCE with CWD and retains the same loss weights on every dataset. SBR selects up to m highest-scoring negatives among the supervised classes. Selection uses detached logits, while gradients flow through the selected positive and negative logits. The ranking term is zero when either set is empty; BCE contributes zero when no valid classes remain, and CWD contributes zero when the total confidence weight is zero. An explicit background category is excluded from presence supervision.

MLLM supervision and shared CWD settings. Qwen3.8-27B generates category-presence scores offline using the prompt described in Sec. 3.3. Condensed prompt templates and example answers with presence probabilities for ADE20K, Cityscapes, and iSAID are provided in Appendix F. The scores are cached for training, so no MLLM is required at segmentation inference. We also keep all CWD hyperparameters fixed across datasets: $h = 0.6$, $\ell = 0.1$, $\epsilon_{\text{pos}} = 0.05$, and $\epsilon_{\text{neg}} = 0.1$. These choices are guided primarily by prior knowledge: a relatively high h favors confident positive labels, while a low ℓ selects reliable negative classes for SBR. Small smoothing rates soften the binary targets while preserving the distinction between positive and negative supervision. With these shared settings, ***TAG-V*** improves all nine main-benchmark metrics over frozen SAM3, demonstrating the effectiveness of a common supervision recipe across domains. The corresponding distance scale in Eq. 10 is $\beta = (h - \ell)/2 = 0.25$. All reported CWD runs use the linear, clipped confidence weight centered at h. Only classes with $p_k \le \ell$ or $p_k \ge h$ enter SBR; classes within the uncertainty band can still contribute to the confidence-weighted BCE.

Image preparation. We use canonical class names as text prompts and reuse their frozen encodings across images. Image features are computed once per input view and shared across class prompts. SAM3 resizes each view to 1008×1008 and normalizes RGB channels with mean and standard deviation both equal to 0.5.

Pooling and initialization. Let $\mu_{tk} \in \{0, 1\}$ indicate whether token t of class k is valid. The pooled anchor used in Eq. 1 is

$$\bar{\mathbf{e}}_k = \frac{\sum_{t=1}^{N_t} \mu_{tk} \, (\mathbf{F}_t^k)_{t,:}}{\max\left(\sum_{t=1}^{N_t} \mu_{tk}, 1\right)}. \tag{11}$$

Each token row $(\mathbf{F}_t^k)_{t,:}$ lies in $\mathbb{R}^{1\times d}$. Applying the same mask to $\mathbf{F}_t'^k(I)$ yields $\bar{\mathbf{e}}'_k$, which conditions the presence and K/V adapters. For ARP, low-rank input projections and the shared presence directions are initialized from a zero-mean Gaussian with standard deviation 10^{-3}, and the image-affinity scale η starts at one. Table 6 lists the zero-initialized output projections, which make every residual vanish at initialization. The gate's final layer also starts at zero, giving a multiplicative gate of one in Eq. 1.

Table 6: **ARP inputs and zero initialization.** Each residual starts at zero through its output projection. K/V updates are applied before splitting features into attention heads.

Module	Residual input	Conditioning feature	Zero-initialized component
Text low-rank branch	$\mathbf{F}_t^k$	$\bar{\mathbf{e}}_k$	output projection $\mathbf{B}$
Text image branch	$\mathbf{F}_t^k, \mathbf{F}_v$	image features and text affinity	Proj_{out} weight/bias
Presence token	$\bar{\mathbf{e}}'_k$	$\bar{\mathbf{e}}'_k$	final layer of MLP_{p}
Fusion text K/V	$\mathbf{F}_t'^k(I)$	$\bar{\mathbf{e}}'_k$	both low-rank output projections
Decoder image K/V	$\mathbf{F}_v^k$	$\bar{\mathbf{e}}'_k$	both low-rank output projections

C.2. Targeted Adaptation Architectures

Figure 6 compares alternative adaptation architectures for ADE20K panoptic segmentation. All trainable variants use image-level category-presence labels and the same BCE and SBR objectives. The pretrained SAM3 weights remain frozen, and only the parameters introduced by each adaptation strategy are optimized. Frozen SAM3 provides the unadapted reference.

Ordinary LoRA. We replace ARP with conventional LoRA (Hu et al., 2022) at two sets of cross-attention projections. In the fusion encoder, LoRA adapts the *textual key and value projections* used by image-to-text cross-attention. In the decoder, it adapts the *visual key and value projections* used by query-to-image cross-attention.

For one such projection, let $\mathbf{X}$ denote its input tokens and $\mathbf{W} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ its frozen pretrained weight, where d_{in} and d_{out} are the input and output widths. We introduce trainable factors $\mathbf{A}_{\text{L}} \in \mathbb{R}^{d_{\text{in}} \times r_{\text{L}}}$ and $\mathbf{B}_{\text{L}} \in \mathbb{R}^{r_{\text{L}} \times d_{\text{out}}}$, where r_{L} is the LoRA rank. The adapted projection is

$$\mathbf{X}\mathbf{W} + (\mathbf{X}\mathbf{A}_{\text{L}})\mathbf{B}_{\text{L}}. \tag{12}$$

Only these factors are optimized; the original projection weights and biases remain frozen. The factors are shared across categories and applied without an explicit class-conditioned gate.

No low-rank updates are added to self-attention, query projections, or attention-output projections. The image and text encoders, feed-forward and normalization layers, prediction heads, and pretrained presence-token embedding also remain frozen. The baseline omits TAG's text-feature adapter, image-conditioned text residual, and class-conditioned presence-token initialization module. The rank allocation is chosen to give a trainable parameter budget comparable to that of the complete ***TAG-I*** model.

Text residual only. This variant confines the added adaptation modules to the text-feature correction in Eq. 3. It retains both the class-conditioned static residual and the image-conditioned text-to-image cross-attention branch, allowing the text anchors to incorporate target-domain semantics and image-specific evidence. The corrected text tokens then enter SAM3's pretrained conditioning pathway. The variant omits the additional fusion and decoder K/V residuals and the class-conditioned presence-token initialization.

Complete *TAG-I*. The full model retains this text-feature correction and adds the remaining ARP components described in §3.1. These components adapt the fusion encoder's text keys and values, initialize the presence token from the corrected class anchor, and condition the decoder's image K/V residuals on that anchor.

***TAG-I* & image residual.** This variant retains the complete ARP and adds a residual branch directly to the image tokens produced by the frozen image encoder and its feature neck, before text-conditioned fusion. For the visual tokens $\mathbf{F}_v$, a learned low-rank mapping Δ_{vis} produces corrected image tokens

$$\widetilde{\mathbf{F}}_v = \mathbf{F}_v + \Delta_{\text{vis}}(\mathbf{F}_v), \tag{13}$$

where both $\widetilde{\mathbf{F}}_v$ and $\Delta_{\text{vis}}(\mathbf{F}_v)$ have the same shape as $\mathbf{F}_v$. The mapping is shared across spatial positions and class prompts, and its output projection is initialized to zero. The corrected visual representation is supplied to both the image-conditioned text branch and the fusion encoder, whose output subsequently feeds the decoder. The added branch is trained jointly with ARP under the same presence objective. In Eq. 2, $\Delta^k_{\text{img}}(I)$ uses visual evidence to update *text tokens*; here, Δ_{vis} directly updates *image tokens*.

D. More Quantitative Experiments

We extend the quantitative evaluation with additional ablation studies, comparisons on VOC and COCO, direct comparisons with SAM3 calibration methods, a comparison of learned conditioning and retrieved class descriptions, object-level recognition and spatial diagnostics, and vocabulary-expansion results on ADE847.

D.1. Ablation Studies

We extend the ablations in Sec. 4.2 to examine the internal conditioning pathway, the choice of MLLM supervision, and the sensitivity to objective weights and adapter capacity.

D.1.1. Module Design

Beyond presence-score correction. Figure 7 compares score calibration and presence classifiers with ***TAG-I*** on ADE20K. All pretrained SAM3 weights remain frozen; only the added modules are trained.

Frozen SAM3. The original model is used without additional modules or adaptation.

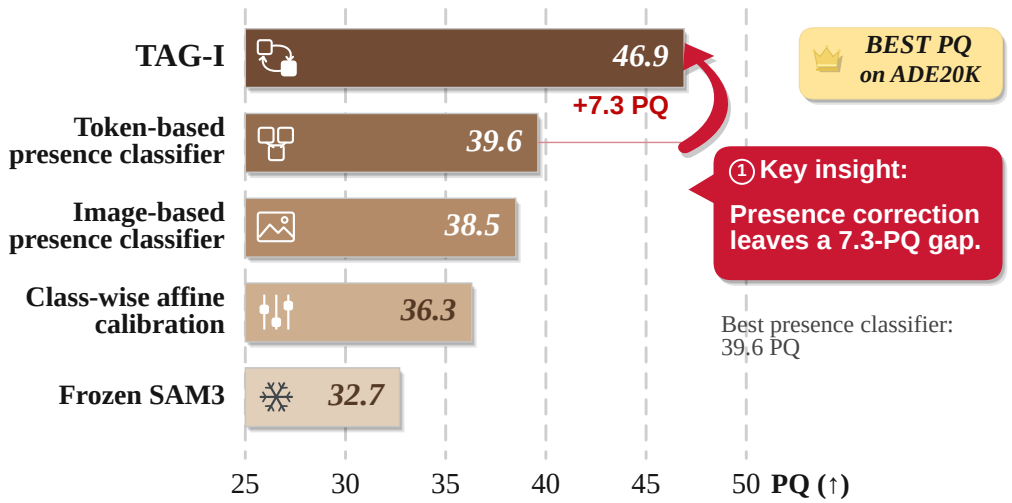

Figure 7: **Presence correction on ADE20K.** PQ of score calibration, presence classifiers, and ***TAG-I***; the axis starts at 25.

Class-wise affine calibration. Only a per-category scale and offset are learned to recalibrate the original presence logits.

Image-based presence classifier. A small classifier corrects the presence logits using pooled features from the frozen image encoder. Only this classifier is trained.

Token-based presence classifier. A small classifier combines the frozen decoder's presence-token features with the queried class's text features to correct the presence logits. Only the classifier is trained; token initialization and the decoder remain unchanged.

***TAG-I*.** ARP corrects text features and adds class-conditioned residuals to the fusion encoder, decoder, and presence-token initialization.

The three output-level controls preserve SAM3's raw mask predictions and change only the presence scores used to filter and combine them.

All three controls improve PQ over frozen SAM3. The token-based classifier reaches 39.6 PQ, the highest score among these controls, and adapting the conditioning pathway with ***TAG-I*** provides a further 7.3 points. These results show that correcting presence scores on fixed features is useful, while adapting how text features guide segmentation provides further gains.

D.1.2. Text Correction and MLLM Supervision

Complementary text and image-conditioned corrections.

Table 7: **Complementary text corrections.** Branch ablations of ***TAG-I***; PQ (%).

Method	ADE	City	iSAID	Avg.	Δ
w/o text low-rank	43.51	59.73	41.39	48.21	−3.05
w/o image conditioning	44.19	61.02	42.60	49.27	−1.99
TAG-I	**46.86**	**62.67**	**44.24**	**51.26**	–

ADE: ADE20K; City: Cityscapes. Avg.: mean over the three datasets; Δ: difference from full ***TAG-I*** in PQ points.

We ablate the two branches that update the text features in Eq. 3. As shown in Table 7, removing the text low-rank branch reduces average PQ from 51.26 to 48.21, while removing image conditioning lowers it to 49.27. Retaining both branches gives higher PQ on all three datasets, supporting their complementary roles. The image-invariant text correction learns category-dependent adjustments that account for the shift from pretrained semantics to the target-domain distribution. The image-conditioned branch injects visual evidence from the current image into the text anchors, making their correction responsive to the observed scene. Together, they combine target-domain semantic correction with image-specific grounding.

MLLM choice and confidence-weighted distillation.

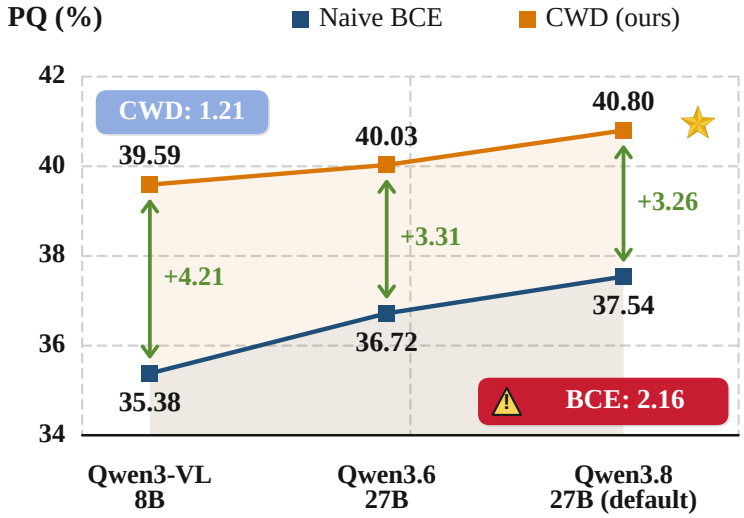

Figure 8: **MLLM supervision on ADE20K.** Naive BCE versus CWD; PQ (%). Green arrows show CWD gains; spans are the ranges across MLLMs.

We focus on open-source MLLMs for offline image-level tagging, avoiding paid API calls. Figure 8 compares naive BCE with CWD on ADE20K. CWD improves PQ by 4.21, 3.31, and 3.26 points with Qwen3-VL 8B, Qwen3.6 27B, and the default Qwen3.8 27B, respectively. Its cross-MLLM performance range is 1.21 points, compared with 2.16 for BCE, indicating lower sensitivity to the tag generator. CWD downweights decisions near the confidence threshold and smooths targets; the accompanying ranking term also excludes uncertain classes. The consistent gains across all three teachers demonstrate that CWD makes effective use of automatically generated presence labels. The default Qwen3.8 27B supervision gives the highest PQ among the tested choices, supporting its use for adapting ***TAG*** on unlabeled target images.

D.1.3. Hyperparameter Sensitivity

A shared configuration across domains. All main-benchmark ***TAG*** results use the same training configuration in Table 5, without dataset-specific tuning. For ***TAG-I***, the common setting $(\lambda_{\text{pres}}, \lambda_{\text{rank}}, m, r, G) = (4, 0.5, 3, 16, 16)$ achieves 46.9, 62.7, and 44.2 PQ on ADE20K, Cityscapes, and iSAID, respectively, as reported in Table 1. The consistent gains over frozen SAM3 across these different domains support the generality of ***TAG*** under a shared training recipe.

Sensitivity protocol. To assess sensitivity to individual design choices, we vary the presence weight λ_{pres}, ranking weight λ_{rank}, hard-negative count m, adapter rank r, and number of shared presence directions G one at a time. All remaining parameters retain the shared configuration above, including the learning rate of 5×10^{-3} and the training budget of 10 epochs. Figures 9 and 10 report absolute PQ on the three datasets, with the common default highlighted in each sweep.

Stable performance across objective weights. Figure 9 varies λ_{pres} from 2 to 6 and λ_{rank} from 0.3 to 0.7. The largest within-dataset PQ spans are 1.28 and 0.91 points, respectively, and the common defaults remain within 0.29 PQ of every sweep's best result. The shared $\lambda_{\text{pres}} = 4$ setting improves PQ over $\lambda_{\text{pres}} = 2$ by 1.06, 0.92, and 0.67 points on ADE20K, Cityscapes, and iSAID, respectively. This comparison supports assigning sufficient weight to presence calibration alongside SBR in Eq. 8. The small default-to-best gaps show that the shared loss weights provide consistently strong performance throughout the evaluated domains.

Stable performance across mining and capacity choices. Figure 10 varies m from 1 to 5 and both r and G from 8 to 24. The maximum PQ spans are 1.07, 1.27, and 1.04 points for m, r, and G, respectively. The default $r = 16$ improves PQ over $r = 8$ by 0.61, 0.89, and 1.00 points on ADE20K, Cityscapes, and iSAID, respectively. Across all mining and capacity sweeps, the common defaults remain within 0.27 PQ of the best result. This combination of compact adapter capacity and near-best accuracy supports the shared choices of $m = 3$, $r = 16$, and $G = 16$.

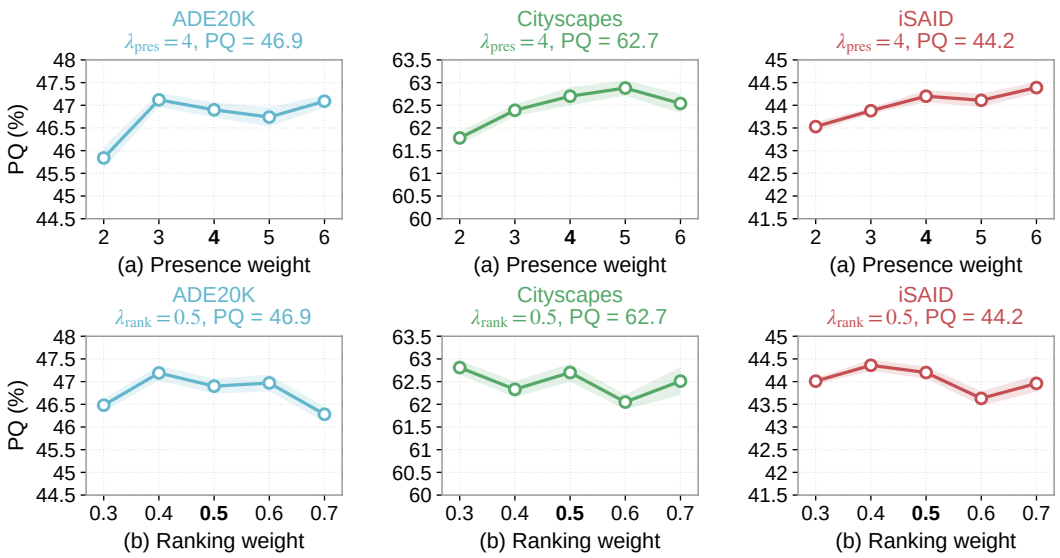

Figure 9: **Sensitivity to objective weights.** Rows vary (a) the presence weight and (b) the ranking weight while fixing all other settings to Table 5. Columns show ADE20K, Cityscapes, and iSAID on separate absolute-PQ axes, each spanning 3.5 points. Axis limits are fixed for each dataset across all sweeps. Colored subtitles report the common setting and its PQ; bold x-axis ticks identify that setting.

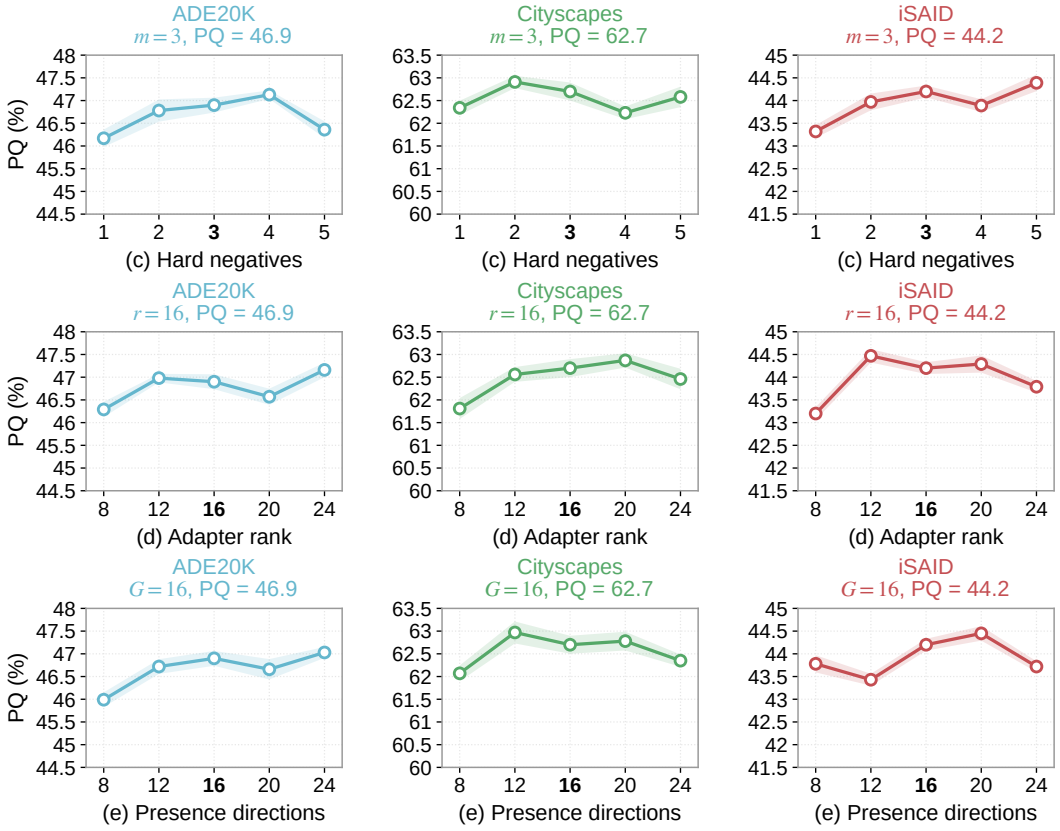

Figure 10: **Sensitivity to hard-negative mining and adapter capacity.** Rows show one-factor sweeps of (c) hard-negative count m, (d) adapter rank r, and (e) shared presence directions G. The common defaults are $m = 3$, $r = 16$, and $G = 16$, with $\lambda_{\text{pres}} = 4$ and $\lambda_{\text{rank}} = 0.5$ fixed. Dataset columns, colors, default annotations, and the 3.5-point axis spans follow Figure 9.

Shared settings across domains. Across all fifteen dataset–parameter pairs, the maximum–minimum difference in central PQ ranges from 0.68 to 1.28 points, while the shared configuration remains within 0.29 PQ of every sweep's best result. These bounded variations support a stable operating range for ***TAG-I*** under individual hyperparameter changes. We therefore use $(4, 0.5, 3, 16, 16)$ on every dataset and keep the CWD settings fixed for ***TAG-V***, as specified in Appendix C.1. The resulting gains of 14.2, 12.4, and 25.0 PQ points over frozen SAM3 on ADE20K, Cityscapes, and iSAID demonstrate effective adaptation across natural images, urban scenes, and aerial imagery with one training recipe.

D.2. Experiments on VOC and COCO

D.2.1. Benchmarks and Metric Conventions

We complement the three challenging benchmarks in Table 1 with the VOC and COCO settings commonly used in weakly supervised segmentation. VOC semantic evaluation uses the 1,449 validation images and 21 classes, including background. ***TAG-I*** uses the 1,464-image VOC training split; the quoted semantic methods use the augmented 10,582-image set. Instance results use the 20 foreground categories. We report both mask AP averaged over IoU thresholds 0.50:0.05:0.95 and AP_{50}, since VOC weakly supervised instance methods commonly report the latter.

COCO evaluation protocols. The semantic methods in Table 9 use *COCO-Object*: 80 thing classes plus a background class. Published semantic results use COCO 2014; our evaluation uses the 118,287/5,000-image COCO 2017 train/validation split, with the two versions reported in separate columns. The panoptic comparison with TRACE and point-supervised methods uses *COCO panoptic 2017*, with 80 thing and 53 stuff classes. For ***TAG-I***, COCO-Object supplies image-level training labels for object-semantic and instance evaluation, and the 171-category COCO-Stuff vocabulary supplies image-level training labels for the COCO-133 panoptic evaluation.

Supplementary training settings. VOC, COCO-Object, and COCO-Stuff use the same training configuration as the main benchmarks in Table 5. The peak learning rate is 5×10^{-3}, the training budget is 10 epochs, and $(\lambda_{\text{pres}}, \lambda_{\text{rank}}, m, r, G) = (4, 0.5, 3, 16, 16)$, without dataset-specific tuning.

D.2.2. Segmentation Gains on VOC and COCO

Table 8 compares ***TAG-I*** with frozen SAM3 on the complete VOC and COCO-Object validation sets. Both instance evaluations retain at most 100 detections per image and use predicted category-presence scores, without GT-present class lists.

Table 8: **SAM3 and *TAG-I* results on VOC and COCO-Object.** AP denotes mask average precision over IoU thresholds 0.50:0.05:0.95; AP_{50} uses IoU 0.50. VOC PQ uses 20 thing classes plus background. COCO PQ uses *81 classes*, merging all stuff into background; Table 10 reports the 133-class evaluation. Bold marks the higher score within each protocol.

Method	mIoU	PQ	AP	AP_{50}
VOC 2012 (1,449 images)				
Frozen SAM3	83.49	81.38	57.34	80.92
TAG-I (Ours)	**89.46**	**85.94**	**59.58**	**83.61**
COCO-Object 2017 (5,000 images)				
Frozen SAM3	63.49	57.62	43.85	68.26
TAG-I (Ours)	**67.58**	**61.22**	**46.56**	**71.03**

The COCO PQ evaluation in Table 8 merges the 53 stuff categories into a single background class while preserving the 80 thing classes and ignored pixels. The 57.62 PQ for frozen SAM3 and 61.22 PQ for ***TAG-I*** are therefore *81-class diagnostic results*. We report the standard 133-class evaluation separately in Table 10. ***TAG-I*** improves every metric in Table 8. On VOC, it gains 5.97 mIoU, 4.56 PQ, 2.24 AP, and 2.69 AP_{50} points over frozen SAM3; the corresponding COCO-Object gains are 4.09, 3.60, 2.71, and 2.77 points. These consistent improvements demonstrate that image-level adaptation strengthens semantic, panoptic, and instance prediction on both benchmarks.

D.2.3. Comparison with Weakly Supervised Methods

Table 9 adds WeCLIP (Zhang et al., 2024), S2C (Kweon & Yoon, 2024), and the CLIP/SAM-assisted framework of Yang & Gong (2024). S2C transfers SAM knowledge into class activation

maps before training a semantic segmenter; the latter framework learns CLIP prompts, uses SAM-assisted seeds, and trains a separate segmentation network. Table 10 additionally compares image-level panoptic methods JTSM and DHR+TRACE, point-supervised PSPS/Point2Mask/EPLD, and instance methods BESTIE, CIM, WeakSAM, and WISH. For WISH, we quote the image-tag-only setting. BESTIE and CIM use their Mask R-CNN refinement results. TRACE's quoted panoptic results combine its diffusion-derived instance edges with DHR semantics.

Table 9: **Image-level semantic segmentation comparisons (mIoU).** Prior denotes the foundation model used by the pipeline; backbone denotes the segmentation backbone. All adapted methods use image-level target labels. The COCO-Object columns contain 80 foreground classes plus background; COCO 2014 and 2017 are reported separately by evaluation split. VOC training-set sizes are specified in the text. Bold marks the highest score in each column.

Method	Prior	Backbone	VOC 2012 val	COCO 2014 val	COCO 2017 val
WeCLIP (Zhang et al., 2024)	CLIP	ViT-B/16	76.4	47.1	–
S2C (Kweon & Yoon, 2024)	SAM	WRN-38	78.2	49.8	–
Yang et al. (Yang & Gong, 2024)	CLIP + SAM	R101	77.3	48.6	–
Yang et al. (Yang & Gong, 2024)	CLIP + SAM	Swin-L	82.6	**55.4**	–
Frozen SAM3	SAM3	SAM3	83.49	–	63.49
TAG-I (Ours)	SAM3	SAM3	**89.46**	–	**67.58**

Table 10: **Weakly supervised panoptic and instance segmentation.** PQ follows VOC-panoptic (20 things plus background) and COCO-panoptic 2017 (133 classes). Instance columns report VOC AP_{50} and COCO 2017 mask AP@[.5:.95]. Sup. follows Table 1; prior denotes the foundation model used. A dash means not reported. Our PQ results use complete validation sets and the same inference procedure for frozen and adapted SAM3. Published methods retain their original training and evaluation settings. Bold marks the highest score in each metric column.

				Panoptic PQ		Instance precision	
Method	Sup.	Prior	Backbone	VOC	COCO-133	VOC AP_{50}	COCO AP
Image-level panoptic methods							
JTSM (Shen et al., 2021)	$\mathcal{I}$	–	R18-WS	39.0	5.3	44.2	6.1
DHR+TRACE (Jo et al., 2026)	$\mathcal{I}$	SD	R50	56.9	32.8	–	–
DHR+TRACE (Jo et al., 2026)	$\mathcal{I}$	SD	Swin-L	69.8	43.1	–	–
Point-supervised panoptic methods							
PSPS (Fan et al., 2022)	$\mathcal{P}$	–	R50	49.8	29.3	–	–
Point2Mask (Li et al., 2023b)	$\mathcal{P}$	–	Swin-L	61.0	37.0	–	–
EPLD (Li et al., 2024b)	$\mathcal{P}$	–	Swin-L	68.5	41.0	–	–
Image-level instance methods							
BESTIE (Kim et al., 2022)	$\mathcal{I}$	–	HRNet-48	–	–	51.0	14.3
CIM (Li et al., 2023c)	$\mathcal{I}$	–	R50	–	–	55.9	17.0
WeakSAM (Zhu et al., 2024)	$\mathcal{I}$	SAM	R50 + M2F	–	–	64.4	25.2
WISH (Kweon & Yoon, 2025)	$\mathcal{I}$	SAM	R50	–	–	62.9	20.7
SAM3 baseline and adaptation							
Frozen SAM3	✗	SAM3	SAM3	81.38	49.18	80.92	43.85
TAG-I (Ours)	$\mathcal{I}$	SAM3	SAM3	**85.94**	**51.57**	**83.61**	**46.56**

Comparison with SAM-based pipelines. On VOC, ***TAG-I*** obtains 89.46 mIoU, exceeding S2C by 11.26 points and the strongest quoted CLIP/SAM-assisted configuration by 6.86 points. Its 83.61 AP_{50} exceeds WeakSAM and image-tag-only WISH by 19.21 and 20.71 points, respectively. On COCO 2017, its 46.56 mask AP exceeds their 25.2 and 20.7 AP by 21.36 and 25.86 points. The higher semantic and instance scores demonstrate the effectiveness of adapting SAM3's text-conditioned recognition with image-level supervision. Among the compared image-level pipelines, ***TAG-I*** is the higher-accuracy choice for VOC semantic segmentation and for instance segmentation on both VOC and COCO.

Panoptic performance and comparison with TRACE. ***TAG-I*** achieves the highest panoptic scores in Table 10: 85.94 PQ on VOC and 51.57 PQ on COCO-133. It exceeds the strongest DHR+TRACE configuration by 16.14 and 8.47 PQ points, respectively, under image-level supervi-

sion. It also surpasses all listed point-supervised methods on both benchmarks, including EPLD by 17.44 and 10.57 PQ points. With matched panoptic inference, the gains over frozen SAM3 are 4.56 and 2.39 PQ points, demonstrating the benefit of ***TAG*** adaptation on the shared foundation model. These results establish ***TAG-I*** as a higher-accuracy option among the evaluated weakly supervised panoptic methods and support reusing pretrained spatial priors through image-level recognition calibration.

D.3. Comparison with SAM3 Calibration Methods

Benchmarks and counterparts. We compare ***TAG*** with ConceptBank (Pei et al., 2026), CoCo-SAM3 (Chen et al., 2026), ActiveSAM (Tien & Shen, 2026), and ProC-SAM3 (Song et al., 2026) in Table 11. We report mIoU on ADE20K, Cityscapes, iSAID, and Pascal Context-59 (PC59) (Mottaghi et al., 2014); PC59 evaluates 59 categories, excluding background. These methods calibrate SAM3's prompts, class selection, or prediction scores for inference without gradient-based training. Their supervision requirements nevertheless differ: *ConceptBank uses ground-truth masks from the target training split* to construct its concept bank offline through mask-pooled visual prototypes, providing spatial supervision unavailable to either ***TAG*** variant.

Result sources and evaluation protocols. ConceptBank's scores are from Tables 1–2 of Pei et al. (2026); CoCo-SAM3 and ActiveSAM's natural-image scores are from Table 1 of their respective papers. ProC-SAM3 and the marked ActiveSAM iSAID scores come from Table I of Song et al. (2026); the latter is their reproduction. The three main-benchmark ***TAG*** scores reuse Table 1. Each quoted method retains its published prompt construction and inference settings. Unreported entries are marked with a dash.

Table 11: **Comparison with SAM3 calibration methods (mIoU, %).** PC59: Pascal Context-59. Target labels specify the supervision used by each method; ConceptBank uses training-set GT masks. Published results retain their original inference protocols. *: ActiveSAM reproduced by Song et al. (2026); –: not reported. **Bold** and underline mark the best and second-best scores in each column. ↑ gives each ***TAG*** variant's gain in mIoU points over the best non-***TAG*** result in that column.

Method	Target labels	ADE20K	Cityscapes	iSAID	PC59
ConceptBank	GT masks	43.3	75.1	35.4	63.0
CoCo-SAM3	None	38.3	70.7	–	61.2
ActiveSAM	None	40.0	69.7	28.3*	60.1
ProC-SAM3	None	–	–	29.5	–
TAG-V (Ours)	MLLM labels	<u>47.0</u> ↑3.7	<u>78.1</u> ↑3.0	<u>57.4</u> ↑22.0	<u>65.2</u> ↑2.2
TAG-I (Ours)	Image labels	**52.7** ↑9.4	**80.3** ↑5.2	**61.4** ↑26.0	**69.6** ↑6.6

Beyond inference-time calibration. ***TAG-I*** and ***TAG-V*** achieve the best and second-best mIoU, respectively, on all four benchmarks. ***TAG-I*** exceeds the strongest non-***TAG*** result by 9.4, 5.2, 26.0, and 6.6 points on ADE20K, Cityscapes, iSAID, and PC59, respectively, reaching 69.6 mIoU on PC59. More notably, ***TAG-V*** surpasses every available calibration counterpart even when trained on unlabeled target images using only image-level labels generated by an off-the-shelf MLLM. Its corresponding gains are 3.7, 3.0, 22.0, and 2.2 points, despite ConceptBank's access to target masks. On iSAID, ***TAG-V*** also outperforms the remote-sensing-specific ProC-SAM3 by 27.9 points (57.4 vs. 29.5). The consistent lead across all four benchmarks demonstrates the accuracy advantage of ***TAG*** over the evaluated SAM3 calibration methods. For applications with unlabeled target images and a known class vocabulary, the comparison favors ***TAG-V*** over the evaluated inference-time calibration methods when segmentation accuracy is the priority.

D.4. Learned Conditioning versus Retrieved Class Descriptions

Concept expansion enriches the language supplied to a segmentor (Pei et al., 2026; Zhu et al., 2026). We examine the benefit of learning the text-conditioned pathway by comparing ***TAG-I*** with nearest-phrase substitution in frozen SAM3. The evaluation covers six classes from ADE20K, Cityscapes, and iSAID.

Retrieval and evaluation. We retrieve alternative descriptions from a fixed bank of 15,057 phrases using cosine scores between normalized, mean-pooled phrase features and corrected text anchors, averaged over 23 discovery images. Table 12 lists the nearest alternatives after excluding the exact

class name. Each substitution replaces the target class's full text encoding in frozen SAM3, preserving all competing classes and output labels. We evaluate mean target IoU on 12 positive validation images per class, disjoint from discovery; iSAID also uses disjoint source images. For this semantic comparison, ***TAG-I*** uses text, fusion, and presence adaptation.

Table 12: **Nearest alternative phrases.** Exact names excluded; †: original name is top-1.

Dataset	Class name	Nearest alternative phrase
Cityscapes	train†	tram
	rider†	rider in the foreground
ADE20K	table†	table furniture
	cushion†	pillow
iSAID	baseball diamond	baseball diamond next to other objects
	ground track field	ground track field in bright daylight

Consistent gains from learned conditioning. ***TAG-I*** achieves the highest mean target IoU for all six classes in Figure 11a. Relative to the original class names, its gains range from 32.0 to 58.3 IoU points, whereas five of the six retrieved alternatives yield gains below one point. These results demonstrate the advantage of learned text-conditioned adaptation over nearest-phrase substitution across the evaluated categories.

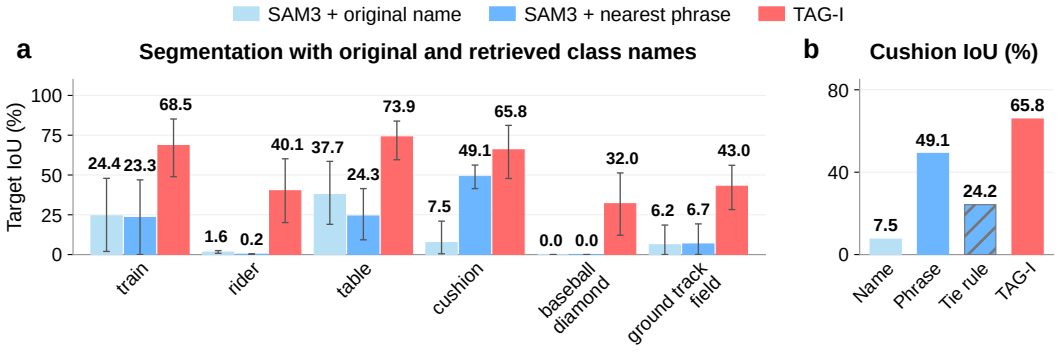

Figure 11: **Segmentation with names, retrieved phrases, and *TAG-I*.** (a) Mean target IoU; whiskers are 95% intervals from 2,000 image-bootstrap resamples. (b) Cushion IoU when exact ties are reassigned to the existing pillow class (hatched), with model scores fixed.

Resolving closely related categories. Cushion is the strongest phrase-substitution result: replacing its prompt with *pillow* raises IoU from 7.5 to 49.1, while ***TAG-I*** reaches 65.8 using the canonical class name. Because *pillow* is also an ADE20K category, the replacement creates identical competing prompts. Assigning exact ties to the existing pillow class gives 24.2 cushion IoU with model scores fixed (Figure 11b). ***TAG-I*** exceeds both substitution settings while retaining distinct category prompts. Together, these comparisons support learning how class anchors condition segmentation to improve target-category recognition.

D.5. Diagnostic Experiments

This section complements the qualitative examples and oracle experiments in Section 2.2. We connect object-level recognition errors, recoverable segmentation accuracy, and the spatial quality of recognized objects to the design of ***TAG***.

D.5.1. Object-Level Error Structure

Diagnostic setup. We analyze frozen SAM3 predictions on ADE20K, Cityscapes, and iSAID. An object is an 8-connected GT foreground component with at least 20 valid pixels; ignored pixels are excluded, and touching instances of the same class may form one component. We regard an object as *failed* when fewer than half of its pixels receive the correct category label. This criterion measures correct-class pixel recall within each GT object. For each failed object, we compute the fraction of its pixels assigned an incorrect label and the fraction assigned its most frequent predicted label, including background where applicable. Both statistics give each failed object equal weight, irrespective of its area. Table 13 records the evaluation populations.

Table 13: Evaluation populations for the object-level spatial statistics. iSAID observations are validation tiles. Failed objects have correct-class pixel recall below 0.5.

Dataset	Images / tiles	GT objects	Failed objects
ADE20K	2,000	40,740	16,281
Cityscapes	500	46,490	19,225
iSAID	11,644	229,354	130,650

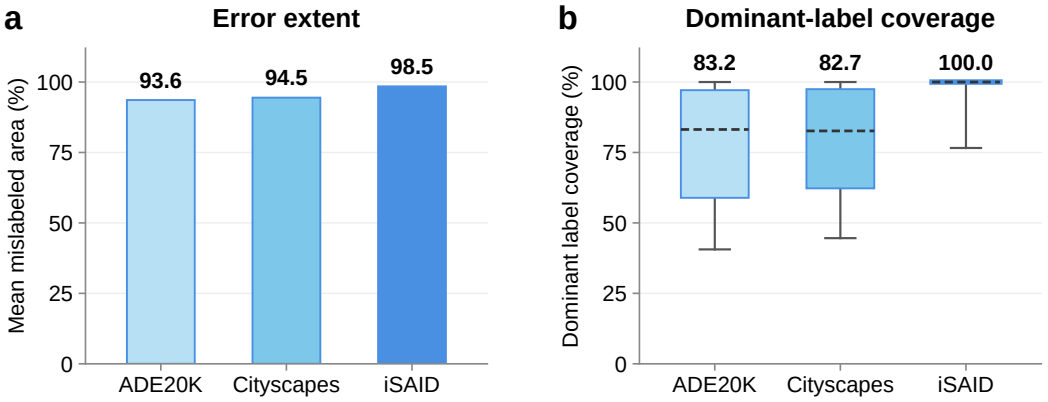

Figure 12: **Object-level structure of frozen SAM3's recognition errors.** Failed objects are GT foreground connected components with correct-class pixel recall below 0.5. **(a)** Mean fraction of mislabeled pixels within each failed object. **(b)** Fraction of each failed object covered by its most frequent predicted label, including background where applicable. Boxes and whiskers denote the interquartile range and 5th–95th percentiles; dashed lines mark medians, and the iSAID box collapses at 100%. Numerical labels give means in (a) and medians in (b). Each failed object contributes equally to both summaries.

Errors are extensive and often dominated by one label. Figure 12a reports mean mislabeled fractions of 93.6%, 94.5%, and 98.5% on ADE20K, Cityscapes, and iSAID, respectively. The high dominant-label medians in Figure 12b indicate that many failed objects receive a consistent wrong category or background label across most of their pixels. The errors therefore frequently extend across the object interior, highlighting category recognition and concept activation as useful targets for adaptation. The oracle interventions below test how much segmentation accuracy can be recovered by correcting these predictions while reusing SAM3's spatial outputs.

D.5.2. Oracle Interventions and Spatial Quality

Oracle definitions. The geometry reference, *GT mask regions*, supplies ground-truth regions while retaining SAM3's category scores. The category and activation interventions in Figure 2b keep SAM3's raw masks, semantic maps, and objectness logits fixed. *GT categories* assigns each 8-connected foreground component of the final prediction its majority GT label over valid pixels, including background, while preserving its pixel support. Ties use total GT class area and then class index; ignored-only components retain their labels. *Positive gates* sets presence scores of GT-present classes to one when scoring and filtering instances and during final fusion, including the explicit background class on iSAID; GT-absent classes retain their original scores. *Gates & categories* applies positive activation followed by majority-GT naming of the resulting foreground components. GT annotations construct these diagnostics; ***TAG*** inference uses class names and predicted presence scores.

Recognition correction recovers segmentation accuracy. As shown in Figure 2b, assigning GT categories to predicted regions improves ADE20K by 47.2 mIoU points over frozen SAM3. Positive activation improves all three benchmarks, and combining category and activation corrections further increases accuracy. These gains demonstrate that better concept recognition can turn existing spatial predictions into substantially more accurate segmentation, directly supporting the motivation for ***TAG***.

Spatial quality of recognized objects. Recognized objects have correct-class recall at least 0.5; both spatial metrics use same-class predictions within the GT component bounding box expanded by 10 px, excluding ignored pixels. Figure 13 reports median boundary F1 at a 5 px tolerance (Perazzi et al., 2016; Cheng et al., 2021a) and median mask IoU for this population. Median mask IoU is 77.9%, 71.7%, and 70.2% on ADE20K, Cityscapes, and iSAID, respectively; median boundary F1 ranges from 76.8% to 87.8%. These measurements show useful boundary and region information for recognized objects across the three domains.

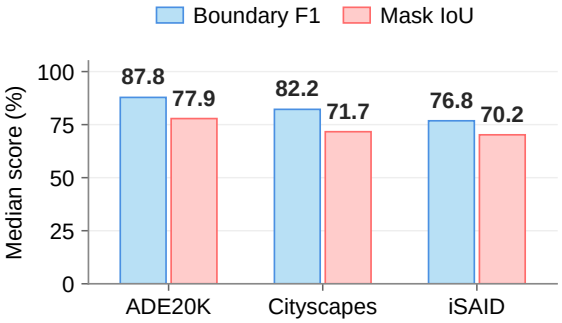

Figure 13: **Spatial quality of recognized objects.** Median boundary F1 at 5 px tolerance and mask IoU (%).

The diagnostics identify recoverable recognition errors alongside useful pretrained spatial information. ***TAG*** builds on these findings by using image-level labels to learn text-conditioned corrections and calibrate presence, while reusing SAM3's pretrained mask heads for dense prediction. The segmentation gains in Table 1 demonstrate the effectiveness of this adaptation strategy across all three domains and tasks.

D.6. New-Vocabulary Evaluation on ADE847

Evaluation protocol. To complement the illustrative categories in Figure 4b, we evaluate ADE150-to-ADE847 transfer on all 2,000 validation images with all 847 labels competing at inference. We transfer the ADE150-trained ***TAG-I*** adapters directly, re-encode the target class names, and use a single-view semantic readout with predicted presence scores. All transferred parameters remain fixed during this evaluation. Ground-truth class lists are never supplied at inference. We accumulate pixel intersections and unions over the validation set before averaging class IoUs. The 150 labels mapped to the training vocabulary and the remaining 697 additional labels are reported separately, while preserving competition among all 847 classes in both groups.

Transferred components. Table 14 compares frozen SAM3, text correction alone, downstream correction alone, and their combination. Text correction applies the trained feature adapter to the newly encoded text; downstream correction reuses the fusion, presence-token, and decoder image-K/V adapters. The downstream-only condition uses the original text features; all conditions share the same vocabulary and semantic readout. The four conditions select which learned adapters are active at inference.

Table 14: **ADE150-to-ADE847 vocabulary expansion (mIoU, %).** All conditions evaluate the complete validation set with 847 candidate labels. Columns partition the labels by correspondence to the ADE150 training vocabulary. **Bold** marks the highest score in each column. The last row gives the combined correction's change from frozen SAM3 in mIoU points.

Transferred components	Mapped 150	Additional 697
None (frozen SAM3)	30.52	11.39
Text correction only	45.18	12.68
Downstream correction only	30.88	12.28
Text + downstream correction	**45.45**	**13.32**
Change from frozen SAM3	+14.93	+1.93

Generalization to additional labels. Combining text and downstream correction improves mIoU on the 697 additional labels from 11.39 to 13.32, a gain of 1.93 points over frozen SAM3. Text correction alone reaches 12.68 mIoU, while downstream correction alone reaches 12.28, corresponding to gains of 1.29 and 0.89 points. Their combination exceeds either component alone, indicating that the two correction pathways contribute complementary benefits to vocabulary expansion. The 150 mapped labels also improve from 30.52 to 45.45 mIoU, a gain of 14.93 points under full-vocabulary competition. The improvements in both groups show that ***TAG*** benefits the expanded ADE vocabulary while retaining strong recognition of the mapped training labels. Together with Figure 4b, the average gain over 697 additional labels demonstrates the utility of reusing learned corrections beyond the training vocabulary.

E. Additional Qualitative Results

Figures 14–16 compare panoptic predictions on eight additional validation scenes per benchmark, illustrating improved category consistency and object coverage after ***TAG*** adaptation.



Figure 14: **Additional panoptic results on Cityscapes.** Columns show the image, ground truth, frozen SAM3, ***TAG-V***, and ***TAG-I***. ***TAG*** improves person/rider and bus/train discrimination and recovers vegetation regions missed by frozen SAM3. All columns use the same crop and category palette.



Figure 15: **Additional panoptic results on ADE20K.** ***TAG*** improves recognition of cushions and furniture categories and recovers regions missed by frozen SAM3, including cradles and kitchen islands. The indoor and outdoor scenes illustrate gains in category consistency and region coverage.

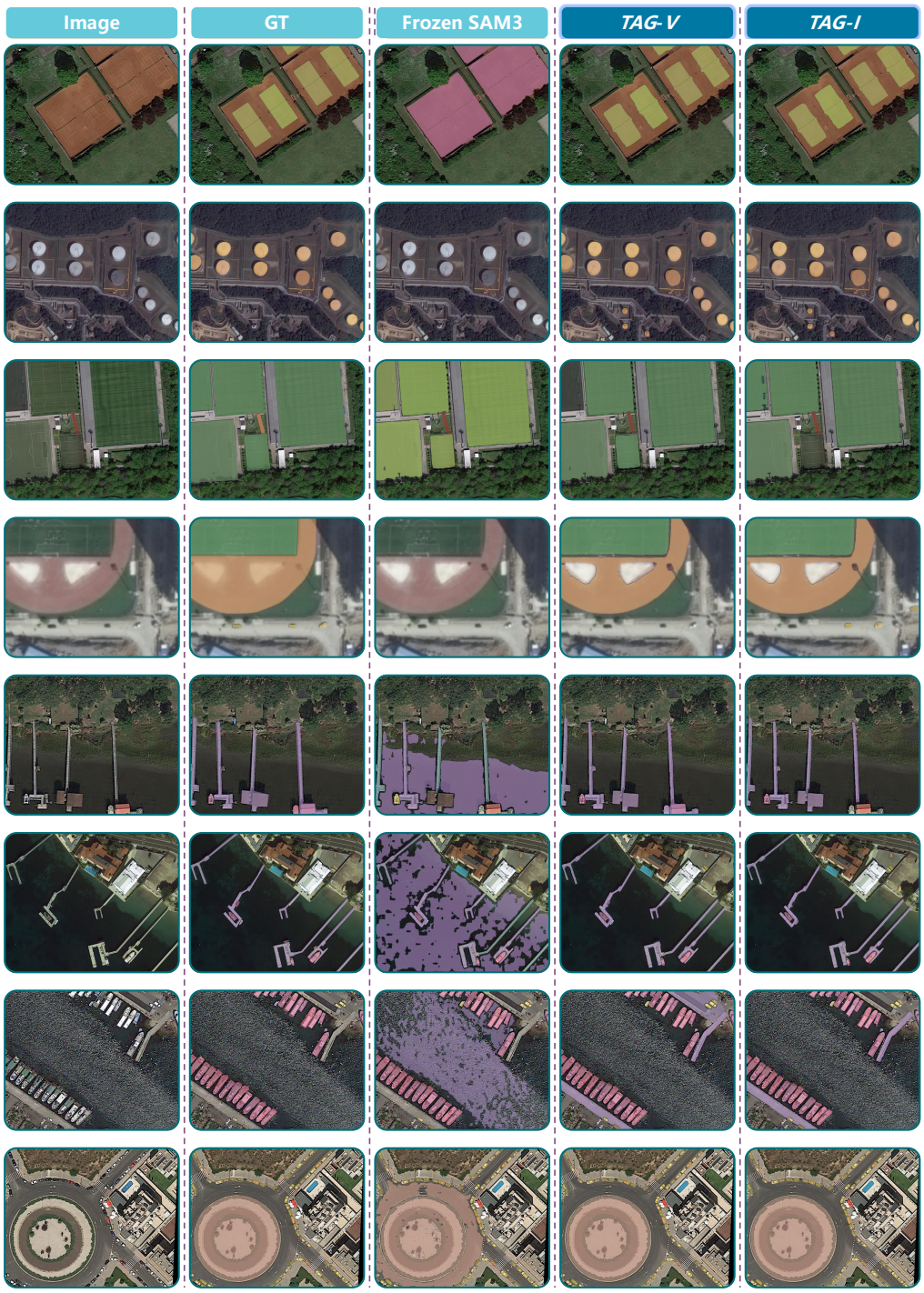

Figure 16: **Additional panoptic results on iSAID.** ***TAG*** improves court/field discrimination, recovers storage tanks and athletics facilities, and reduces foreground leakage around piers. The last two rows show predictions for crowded boats and small vehicles.

F. Dataset-Specific Prompts for Presence Labeling

The following condensed templates retain representative category rules. Each example entry is `[answer, probability]`: a yes/no decision followed by the token-derived probability $p_k \in [0, 1]$ of category presence (Sec. 3.3). CWD converts p_k into training targets and confidence weights using Eqs. 9 and 10. A short description summarizes the input image for each example.

ADE20K: Indoor and Outdoor Scenes

The field `{category_list}` denotes the supplied ADE20K categories; the full vocabulary has 150 classes. Only representative names, answers, and scores are displayed.

ADE20K: condensed presence-labeling prompt

```
This is a photograph of an indoor or outdoor scene. Identify the categories
  visible in the image.

Candidate categories: {category_list}
Examples: wall, floor, ceiling, bed, windowpane, curtain, pillow, lamp, chair,
  sofa, person, ...

Rules:
1. Inspect the entire image, including both objects and visible background
  surfaces.
2. Judge presence from visual evidence. Scene context may help, but a typical
  object need not be present.
3. Include small, distant, partly occluded, or truncated objects when they
  remain recognizable.
4. Distinguish related categories by their visual role; for example, a pillow
  on a bed differs from a cushion on a sofa.
5. Keep the supplied category names unchanged, and consider only categories in
  the candidate list.

Response format:
Return only a JSON object with one entry per supplied category and "yes" or
  "no" as its value.
```

Example input. An RGB bedroom image showing walls, a floor, a ceiling, a bed with pillows, a curtained window, and a bedside lamp. The dictionary below shows selected categories; all supplied categories receive an answer and a score in the full output.

ADE20K: example answers and probabilities (selected classes)

```
{
  "wall": ["yes", 0.99], "floor": ["yes", 0.98],
  "ceiling": ["yes", 0.96], "bed": ["yes", 0.99],
  "windowpane": ["yes", 0.91], "curtain": ["yes", 0.88],
  "pillow": ["yes", 0.92], "lamp": ["yes", 0.93],
  "cushion": ["no", 0.37], "chair": ["no", 0.16],
  "sofa": ["no", 0.08], "person": ["no", 0.02]
}
```

Cityscapes: Urban Street Scenes

The field {category_list} denotes the 19 Cityscapes categories. The template highlights a few common distinctions, while the example includes answers and probabilities for all 19 keys. Canonical names such as trafficlight and trafficsign are preserved.

Cityscapes: condensed presence-labeling prompt

```
This is a street-level photograph taken from a vehicle. Identify the
  categories visible in the image.

Candidate categories: {category_list}
Examples: road, sidewalk, building, vegetation, terrain, person, rider, car,
  bicycle, ...

Rules:
1. Scan the whole scene, including distant objects, image edges, and partly
  occluded instances.
2. Distinguish pedestrians from riders. A visible cyclist and bicycle support
  both "rider" and "bicycle".
3. Separate low grass or exposed ground ("terrain") from trees and bushes
  ("vegetation").
4. Identify vehicles from visible appearance; for example, reserve "bus" for
  recognizable buses rather than uncertain distant vehicles.
5. Ignore the recording vehicle's hood. Use only the supplied categories and
  preserve their exact names.

Response format:
Return only a JSON object with one entry per supplied category and "yes" or
  "no" as its value.
```

Example input. An RGB street image with roads, sidewalks, buildings, trees, a narrow grass strip, traffic signals, cars, a pedestrian, and a cyclist. The example illustrates scores for one view, including uncertainty about a small terrain region.

Cityscapes: example answers and probabilities (19 classes)

```
{
  "road": ["yes", 0.99], "sidewalk": ["yes", 0.97],
  "building": ["yes", 0.98], "wall": ["no", 0.12],
  "fence": ["no", 0.24], "pole": ["yes", 0.94],
  "trafficlight": ["yes", 0.91], "trafficsign": ["yes", 0.89],
  "vegetation": ["yes", 0.95], "terrain": ["yes", 0.58],
  "sky": ["yes", 0.99], "person": ["yes", 0.93],
  "rider": ["yes", 0.87], "car": ["yes", 0.98],
  "truck": ["no", 0.18], "bus": ["no", 0.07],
  "train": ["no", 0.01], "motorcycle": ["no", 0.11],
  "bicycle": ["yes", 0.92]
}
```

Each entry pairs the yes/no answer with its cached category-presence probability. For multiple labeling views, the per-class maximum supplies the image-level score.

iSAID: Overhead Aerial Imagery

The field {category_list} denotes iSAID's 15 foreground classes and background. The template summarizes visual distinctions relevant to aerial imagery; the example includes all 16 keys with their canonical spellings.

iSAID: condensed presence-labeling prompt

```
This is an overhead aerial image tile. Identify the categories visible from
  the top-down view.

Candidate categories: {category_list}
Examples: ship, harbor, store tank, bridge, large vehicle, small vehicle,
  plane, background, ...

Rules:
1. Inspect the full tile for small objects. Recognizable objects cut by the
  tile boundary still count.
2. Require visible object evidence; for example, a runway alone does not
  indicate "plane".
3. Distinguish related structures: ships are vessels, while harbors are dock
  or pier structures. Both may be present.
4. Use size and shape to separate similar categories, such as small versus
  large vehicles or tanks versus roundabouts.
5. Use only the supplied names. Treat roads, buildings, vegetation, and other
  non-target regions as background.

Response format:
Return only a JSON object with one entry per supplied category and "yes" or
  "no" as its value.
```

Example input. An RGB overhead port tile containing ships, piers, circular storage tanks, cars, and elongated trucks. Roads, buildings, and open water form the background. Small vehicles and structures can receive intermediate scores when their appearance is ambiguous.

iSAID: example answers and probabilities (16 classes)

```
{
  "background": ["yes", 1.00], "ship": ["yes", 0.96],
  "store tank": ["yes", 0.88], "baseball diamond": ["no", 0.03],
  "tennis court": ["no", 0.04], "basketball court": ["no", 0.06],
  "ground track field": ["no", 0.02], "bridge": ["no", 0.23],
  "large vehicle": ["yes", 0.68], "small vehicle": ["yes", 0.84],
  "helicopter": ["no", 0.02], "swimming pool": ["no", 0.03],
  "roundabout": ["no", 0.09], "soccer ball field": ["no", 0.02],
  "plane": ["no", 0.04], "harbor": ["yes", 0.98]
}
```

Each entry pairs the yes/no answer with its cached category-presence probability. The background score is fixed to one and excluded from presence supervision (Appendix C.1).