SGQA: Semantic-Geometric Quality Alignment for Training-Free Few-Shot Instance Segmentation

Yuhao Qing*, Liuyan Feng*, and Haoyuan Li**

Shanghai University, Shanghai, China
yuhao_fly@163.com, {fengly2025,lihaoyuan}@shu.edu.cn

Abstract. Compositions of frozen foundation models offer a training-free route to instance segmentation, yet a clear gap remains between current systems and their empirical upper bounds. We introduce *progressive oracle replacement*, a diagnostic procedure that decomposes this gap into stage-level components, and find that the scoring stage accounts for most of it. Across diverse encoders and mask generators, we trace this gap to *Quality Misalignment*: a mismatch between semantic matching confidence and mask geometric fidelity that yields ranking errors when the two signals are used independently. This pattern persists across all tested configurations, suggesting that it is intrinsic to decoupled scoring. We therefore propose Semantic-Geometric Quality Alignment (SGQA), which replaces decoupled scoring with a single quality score formed by the geometric mean of the two signals and adds no hyperparameters to the scoring formula. Under a fixed backbone of DINOv2 ViT-L paired with SAM2 Hiera-L, SGQA outperforms existing training-free methods and matches or exceeds fine-tuning-based methods with comparable or smaller backbones.

Keywords: Few-shot instance segmentation · Foundation model composition · Quality alignment · Training-free

1 Introduction

Large-scale pre-trained foundation models such as SAM [20] and SAM2 [34] for mask generation, and DINOv2 [28] for visual understanding, have been applied across few-shot object detection [51], one-shot segmentation [26], and personalized recognition [49]. Composing these frozen models enables training-free instance segmentation: a semantic encoder identifies *what to look for*, and a mask generator determines *where to segment*.

This composition is, however, not well suited to few-shot instance segmentation. At the scoring stage, semantic quality and geometric quality are treated separately, so their joint quality is not accounted for. As shown in Fig. 1(a), directly composing DINOv2 and SAM2 yields a representative mis-ranking: a

* Equal contribution.
** Corresponding author.

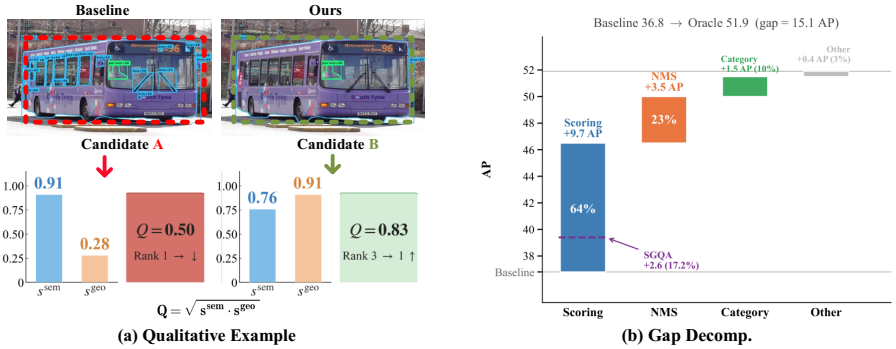

Fig. 1: Quality Misalignment in frozen foundation model composition. **(a)** Decoupled scoring mis-ranks a false positive with high semantic confidence but poor mask quality above a true positive. SGQA aligns both quality signals and corrects the ranking. **(b)** Oracle decomposition of three pipeline stages reveals that scoring is the dominant performance bottleneck.

false positive with a high semantic score but a poor mask is ranked above a true positive whose two quality scores are more balanced.

Recent training-free methods improve few-shot segmentation through feature matching, prompt personalization, or prototype construction [26, 49, 51]. However, they keep semantic and geometric quality signals decoupled, which causes ranking errors when the signals disagree. We therefore ask where information is lost in frozen model composition and how the loss manifests.

To locate the loss, we examine the composition pipeline stage by stage. An oracle decomposition (Fig. 1(b)) shows that the scoring stage alone accounts for most of the performance gap, pointing to signal alignment rather than model capability.

A score-space analysis exposes a consistent distributional pattern: true positives require both quality scores to be high, whereas false positives show a pronounced imbalance between the two (Sec. 3.2). Rank correlation analysis indicates that the two signals carry comparable but individually insufficient ranking information, which motivates an equal-weight fusion. We refer to this pattern as *Quality Misalignment*, the frozen-composition counterpart of the classification-localization misalignment studied in supervised detection [7, 16, 23], and observe it across all tested encoder families (Sec. 3.3).

This diagnosis suggests that part of the performance gap in recent methods [26, 51] stems from unresolved *Quality Misalignment*. It points to two guidelines for frozen model composition: align the two quality signals before ranking and suppression, and weight semantic and geometric information equally. Both lead to a minimal change at the scoring stage, replacing the decoupled scores with the geometric mean of the semantic and geometric quality scores. We call this score-level fusion Semantic-Geometric Quality Alignment (SGQA).

We make two contributions. First, we introduce progressive oracle replacement with factorial analysis for frozen model composition pipelines. Across eight encoders spanning three training paradigms and four mask generator variants, this analysis identifies *Quality Misalignment* as a recurring bottleneck, attributes 64% of the performance gap to the scoring stage, and shows why score-level correction has an inherent ceiling. Second, guided by this diagnosis, we propose SGQA, a geometric-mean fusion that adds no hyperparameters to the scoring formula and recovers 42% of the score-level-recoverable gap (2.6 of 6.2 AP), with consistent improvements across all tested configurations and no additional parameters or training.

2 Related Work

Few-Shot Object Detection and Instance Segmentation. Existing methods fall into three paradigms. **Meta-learning methods** [5,46,47] learn transferable representations through episodic training on base categories. **Fine-tuning methods** [30, 39, 41] pre-train on abundant base classes and adapt to novel categories through controlled parameter updates. **Training-free methods** leverage frozen foundation models without parameter updates: DE-ViT [51] performs prototype matching on frozen DINOv2 features, Matcher [26] combines DINOv2 and SAM [20] via feature correspondences, and PerSAM [49] derives personalized point prompts from a single reference. Training-free compositions now span open-set detection [25], foundation model assembly [36], in-context segmentation [42, 43], open-vocabulary and unsupervised segmentation [31, 32], and visual prompt tuning [15], yet few explicitly ask where information is lost when frozen models are composed. Open-vocabulary grounding detectors such as DINO-X [35] instead rely on supervised pre-training at grounding scale; they are complementary to our training-free, few-reference setting rather than substitutive, and could themselves serve as one of the frozen encoders that SGQA composes.

Quality Estimation in Object Detection. The misalignment between classification confidence and localization quality is a long-studied problem in object detection. IoU-Net [16] first identified this disconnect and added a learned IoU prediction branch for NMS re-ranking. Cascade R-CNN [2] addresses quality through multi-stage refinement. Centerness [40] offers an anchor-free alternative that multiplies a geometric quality signal with the classification score during NMS, which is the same form as our score product, and adaptive sample selection [50] takes a related route. GFL [23] unifies classification and localization quality into a single representation during training, and TOOD [7] and VFNet [48] refine this task alignment with dedicated losses. Soft-NMS [1] addresses a related issue at the suppression stage. SGQA shares the core insight of these works: semantic confidence and spatial quality should be considered jointly when ranking detections. Our contribution is not this insight, but two points specific to the frozen composition setting. First, our progressive oracle replacement (Sec. 3.3) shows quantitatively that this classical misalignment re-emerges as the

dominant bottleneck when frozen foundation models are composed without joint optimization. Guo et al. [10] show that a single network can be poorly calibrated. Our setting is different, in that scores from two *independently* pre-trained models are not designed to be comparable in scale or ranking. We show this cross-model misalignment empirically, in a setting where end-to-end solutions such as GFL do not apply by construction. Second, we show that a geometric mean, supported by ablation, recovers a substantial fraction of the oracle-attainable improvement with no learnable components.

Score fusion is also studied in model ensembling (e.g., Weighted Box Fusion [38]) and multi-modal calibration (e.g., Platt scaling [29]). These methods assume a shared training objective or access to a calibration set, neither of which holds for two independently pre-trained models with no joint optimization, so learned calibration is infeasible. In our comparison, temperature scaling yields only +0.5 AP, confirming that calibration alone is insufficient in this setting.

Diagnostic Analysis in Vision Systems. Oracle analysis has long been used to diagnose detection bottlenecks [14]. Recent work applies similar diagnostics to single-model internals: ClearCLIP [21] traces dense-prediction noise to residual connections, and SegEarth-OV [22] identifies anomalous CLS token responses, each deriving a targeted fix from the analysis. We instead diagnose a *multi-model composition* rather than a single architecture, using progressive oracle replacement to show that scoring accounts for 64% of the gap across all eight tested encoders and four mask generators.

3 Method

3.1 Preliminaries

Following prior work [26, 51], we use "training-free" to denote methods requiring no gradient-based optimization or learnable parameter updates. Non-parametric operations such as K-Means clustering and PCA are permitted under this definition.

Problem Setting. We study few-shot instance segmentation without training. Given a support set $\mathcal{S} = \{(I_j^c, M_j^c)\}_{c=1,j=1}^{C,K}$, where I_j^c is the j-th reference image of class c, M_j^c is the corresponding instance mask, and K is the number of shots per class, the objective is to detect and segment all instances of C categories in a query image I_q without any parameter updates.

Memory Bank Construction. Following [3] (Fig. 2(a,b)), for each category c, a frozen semantic encoder $\mathcal{E}_{\text{sem}}$ (*e.g.*, DINOv2 ViT-L) extracts dense features from K reference images, and the annotated masks filter foreground regions. Cosine-similarity-based K-Means clustering is then applied to the foreground features, yielding cluster centers $\{\mathbf{c}_k^c\} \in \mathbb{R}^D$ that serve as category prototypes for inference.

Compositional Inference Pipeline. As illustrated in Fig. 2(c), the baseline composes two frozen models for candidate-level segmentation and classification. Given a query image I_q, the mask generator $\mathcal{G}_{\text{mask}}$ (*e.g.*, SAM2) produces N

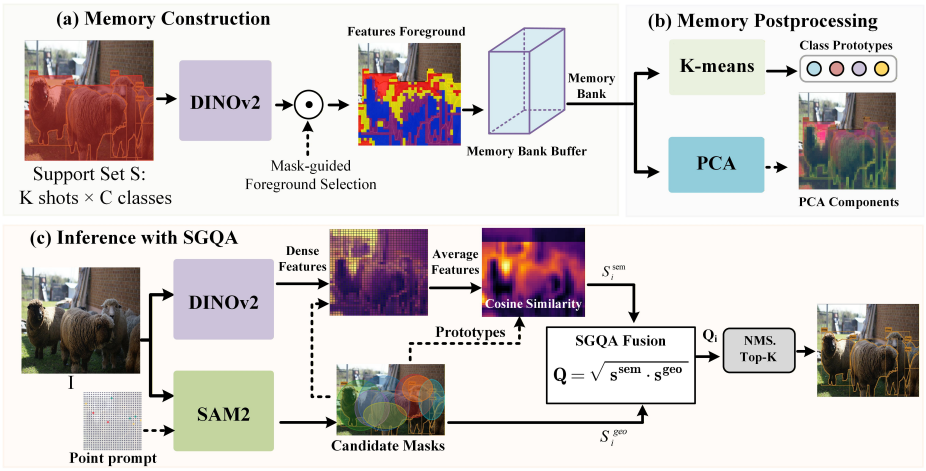

Fig. 2: Overview of the SGQA framework. (a) Memory construction: a frozen semantic encoder extracts dense features from K reference images per class. **(b)** Memory postprocessing: K-Means clustering yields compact class prototypes; PCA is computed for diagnostic visualization. **(c)** Inference: the semantic score s_i^{sem} and geometric score s_i^{geo} are fused via geometric mean into a unified quality score for ranking and detection.

candidate masks $\{m_i\}_{i=1}^{N}$ from a uniform 32×32 point grid, each accompanied by a predicted IoU score $s_i^{\text{geo}} \in [0,1]$ reflecting geometric quality. The encoder $\mathcal{E}_{\text{sem}}$ then extracts the average foreground feature $\bar{\mathbf{f}}_i \in \mathbb{R}^D$ for each mask m_i and compares it against the prototypes $\{\mathbf{c}_k^c\}$ of all C categories via cosine similarity, yielding a semantic score $s_i^{\text{sem}} \in [0,1]$ and a category assignment. Candidates are ranked by s_i^{geo} for non-maximum suppression (NMS), and the reported detection confidence is the semantic score

$$s_i^{\text{sem}} = \max_{c \in \{1,\dots,C\}} \max_{k} \cos(\bar{\mathbf{f}}_i,\ \mathbf{c}_k^c) \tag{1}$$

where $\bar{\mathbf{f}}_i = \frac{1}{|m_i|} \sum_{p \in m_i} \mathcal{E}_{\text{sem}}(I_q)_p$ denotes the average feature within the mask region.

3.2 Motivation

The baseline in Eq. (1) treats the two frozen models as independent modules, so the semantic and geometric quality signals are never compared during ranking. Prior training-free methods [26, 49, 51] refine feature matching [26], prompt personalization [49], or prototype construction [51], each improving a single component rather than their interaction.

These methods still leave s^{sem} and s^{geo} decoupled throughout the pipeline. Plotting the joint distribution of true and false positives in the score space exposes a consistent pattern (Fig. 3): true positives cluster in the upper-right region

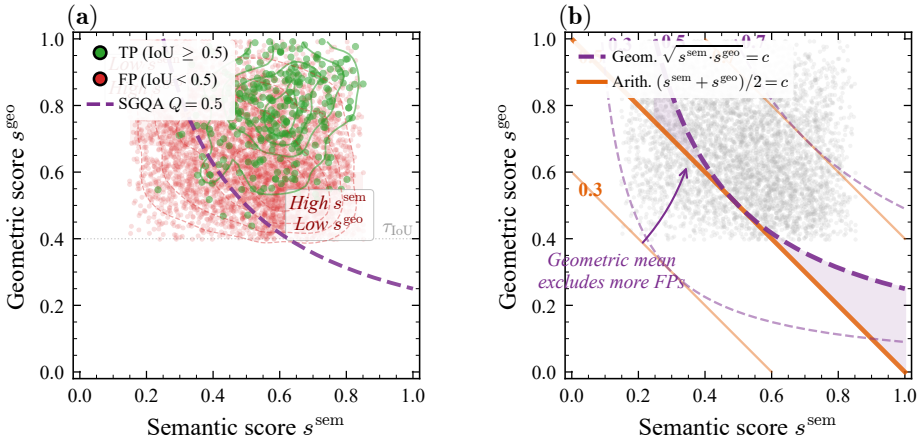

Fig. 3: Score space analysis on COCO 30-shot. (a) Joint distribution of true positives and false positives in the $(s^{\text{sem}}, s^{\text{geo}})$ space with the SGQA decision boundary. True positives cluster in the high-agreement region while false positives scatter along the misalignment axes. **(b)** Iso-value contours of the geometric mean versus the arithmetic mean. The geometric mean penalizes score misalignment more aggressively.

where both scores are high, whereas false positives spread along the off-diagonal regions where one score is high and the other low. Under the baseline, candidates with high s^{sem} but low s^{geo} are thus ranked above true positives with more balanced scores. Refining either component in isolation does not remove this ranking bias, which we next localize within the pipeline.

3.3 Diagnosing Quality Misalignment

The disconnect between classification confidence and localization quality is well recognized in supervised object detection (Sec. 2). We use the term ***Quality Misalignment*** for its form in the frozen composition setting, where the two scores come from independently pre-trained models that were never jointly optimized. This structural independence is what separates our setting from the supervised case, in which end-to-end training can implicitly align the two signals.

Definition 1 (Quality Misalignment). *Given a candidate set* $\{m_i\}_{i=1}^{N}$ *with semantic scores* s_i^{sem} *and geometric scores* s_i^{geo}*, let* $q_i^* = \max_j \text{IoU}(m_i, g_j)$ *denote the true detection quality. The* mis-ranking rate *of a scoring function* $s(\cdot)$ *is*

$$R(s) \;=\; \frac{\left|\{(i,j) : s(m_i) > s(m_j) \;\wedge\; q_i^* < q_j^*\}\right|}{\left|\{(i,j) : q_i^* \neq q_j^*\}\right|}. \tag{2}$$

We say that a compositional pipeline exhibits systematic quality misalignment *if (i) the baseline scoring function* s_0 *has a high mis-ranking rate,* $R(s_0) \gg 0$*, and*

(ii) mis-ranked pairs are concentrated among candidates with large score divergence, i.e., the conditional mis-ranking rate among high-divergence candidates exceeds the unconditional rate: $R(s \mid |s_i^{\text{sem}} - s_i^{\text{geo}}| > \text{median}) > R(s)$.

Operationalization. Note that $R(s) = (1-\tau_b(s, q^*))/2$, where τ_b is the Kendall rank correlation coefficient [19]. Computing $R(s)$ exactly requires enumerating all candidate pairs, which is prohibitive for large pools, so we use two proxies throughout our experiments. The first is the Spearman rank correlation between a scoring function and ground-truth IoU, which captures the global ranking quality that $R(s)$ formalizes. The second is COCO AP, which measures the downstream effect of ranking quality on detection. Definition 1 fixes the concept, and these two metrics provide tractable approximations of the same ranking quality.

The baseline pipeline (Sec. 3.1) makes three decisions: scoring ranks candidates, NMS suppresses redundancy, and label assignment classifies detections. We isolate each decision to measure its contribution to the performance gap.

Oracle Decomposition. To this end, we replace each decision individually with its ground-truth oracle while leaving the others unchanged, which isolates the cost of that decision. We substitute scoring with ground-truth IoU, NMS ranking with ground-truth IoU, and classification with ground-truth labels. Results on COCO 30-shot appear in Fig. 1(b).

Oracle scoring yields a larger improvement than oracle NMS or oracle label and accounts for most of the gap (Fig. 1(b)). SAM2 already reaches a high AR@100, so correct masks are present in the candidate pool but are not ranked highly. The bottleneck is therefore the scoring stage, not the capability of either model.

Combining oracle scoring with oracle NMS recovers 82% of the gap (reaching 49.2 AP), which confirms scoring as the primary target. The goal thus shifts from improving model capability to aligning the two signals produced by the composed models.

To see why decoupled scores rank poorly, we compute the Spearman correlation of s^{sem} and s^{geo} with the ground-truth IoU of the best-matching instance (Fig. 4). Two observations follow. First, ρ_{sem} and ρ_{geo} are moderate and of comparable magnitude, so neither score alone ranks candidates well, while their geometric mean correlates more strongly with IoU. Second, this comparable magnitude motivates symmetric fusion, a choice we later validate by ablation (Sec. 4.2). The same pattern holds when we repeat the analysis with DINOv3 [37], CLIP [33], and the remaining encoders (eight in total), indicating a property of decoupled scoring rather than of a particular encoder: the semantic score does not reflect mask geometry, and the geometric score does not reflect semantics. We refer to this disconnect as *Quality Misalignment*. It also explains why prior training-free methods [26,51] remain subject to the same ranking bias, since they never align the two signals.

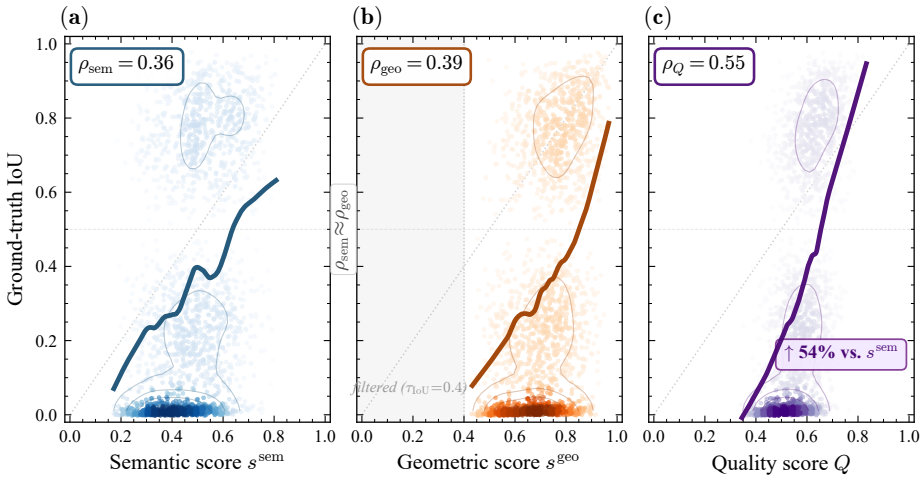

Fig. 4: Spearman rank correlation with ground-truth IoU. (a) Semantic score s^{sem}. **(b)** Geometric score s^{geo}. **(c)** Unified quality score Q. Neither individual score alone provides sufficient ranking quality, while their geometric mean yields substantially higher correlation.

3.4 SGQA: Semantic-Geometric Quality Alignment

Unified Quality Scoring. The diagnosis in Sec. 3.3 identifies *score misalignment* as the primary failure mode: candidates with one high and one low score are ranked above candidates that score well on both. The scoring function should therefore require a candidate to be good on semantic relevance and geometric fidelity at the same time.

We propose to replace the decoupled scoring with a unified quality score based on the weighted power-mean family:

$$Q(m_i) \;=\; (s_i^{\text{sem}})^{\alpha} \cdot (s_i^{\text{geo}})^{1-\alpha}, \quad \alpha \in (0,1). \tag{3}$$

This formulation subsumes several training-free methods as limiting cases: DE-ViT [51] ranks mainly by semantic similarity ($\alpha \to 1$), whereas Matcher [26] and PerSAM [49] rely mostly on the mask generator's IoU head ($\alpha \to 0$). Both extremes inherit the score misalignment diagnosed in Sec. 3.3, which motivates a balanced, intermediate α. Setting $\alpha = 0.5$ yields the geometric mean,

$$Q(m_i) = \sqrt{s_i^{\text{sem}} \cdot s_i^{\text{geo}}}, \tag{4}$$

which we adopt as the default. The comparable magnitudes of ρ_{sem} and ρ_{geo} indicate that neither signal alone is a dominant quality proxy, motivating symmetric fusion. A fine-grained ablation over α (Fig. 5 and Sec. 4.2) confirms that α=0.5 is optimal or near-optimal across all tested configurations. The geometric mean also admits a probabilistic reading: if s^{sem} and s^{geo} are treated as approximately conditionally independent likelihood ratios given candidate quality, the

Bayesian posterior is proportional to their product, which equals the geometric mean up to normalization. Compared with the arithmetic mean, it is more robust to score-scale mismatch, because a single low score keeps the fused score from being dominated by the other. The power-mean comparison in Sec. 4.2 further shows that p=0 outperforms the other tested exponents. We write $Q_i \triangleq Q(m_i)$ for brevity. Q_i serves as both the NMS ranking criterion and the final detection confidence (cf. Fig. 2). SGQA adds no hyperparameters and requires no training. Because the square root is monotone, $\sqrt{s^{\text{sem}} \cdot s^{\text{geo}}}$ and $s^{\text{sem}} \cdot s^{\text{geo}}$ produce identical rankings. The square root only rescales the confidence magnitudes, which matters for the threshold-sensitive COCO precision-recall computation.

4 Experiments

We evaluate SGQA through ablation studies, diagnostic experiments, and comparisons with prior few-shot instance segmentation methods. Section 4.1 describes the setup, Sec. 4.2 reports the ablations and diagnostics on *Quality Misalignment*, and Sec. 4.3 compares against existing methods.

4.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate on two benchmark datasets widely employed for few-shot instance segmentation. The primary benchmark is COCO 2017 [24] with the standard 20-class few-shot split, where $K \in \{10, 30\}$ reference images per class are sampled from the training set and evaluation is performed on the full validation set. For cross-dataset generalization, we additionally evaluate on PASCAL VOC [4] with three base/novel class splits under $K \in \{1, 2, 3, 5, 10\}$ shots. We follow the COCO evaluation protocol and report Bbox AP and Segm AP averaged over IoU thresholds [0.5:0.05:0.95], along with AP_{50}, AP_{75}, and scale-wise $\text{AP}_S/\text{AP}_M/\text{AP}_L$.

Implementation Details. The semantic encoder is DINOv2 ViT-L/14 [28], frozen with input resolution 518×518 and D=1024. The mask generator is SAM2 Hiera-Large [34], also frozen, prompted with a 32×32 uniform point grid at a batch size of 256. Masks are filtered at predicted IoU > 0.4, followed by NMS at IoU > 0.5, retaining up to 100 detections. The memory bank stores 4 K-Means cluster centers per class. As the method requires no retraining or fine-tuning, evaluation proceeds directly on the validation sets. All experiments are conducted on four NVIDIA RTX 4090 GPUs with PyTorch 2.3.1.

Baseline Methods. We compare our method against three categories of few-shot instance segmentation approaches: 1) **Meta-learning methods**, including Meta R-CNN [46] and SNIDA [44]; 2) **Fine-tuning methods**, including TFA [41] and DeFRCN [30]; 3) **Training-free methods**, including DE-ViT [51]. NTT [3] (No Time to Train) refers to the DINOv2+SAM2 composition pipeline without SGQA, as described in Sec. 3.1.

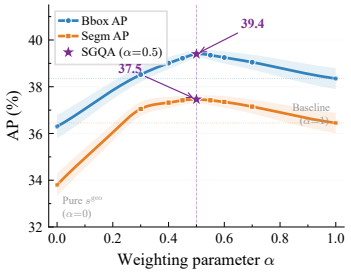

Fig. 5: Sensitivity to the weighting parameter α on COCO 30-shot. Performance peaks at equal weighting and degrades on both sides.

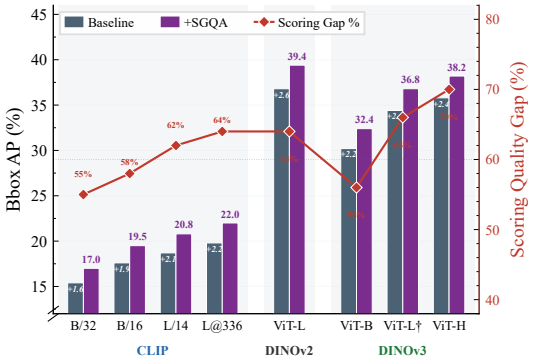

Fig. 6: Cross-encoder consistency on COCO 30-shot. Grouped bars: baseline (*dark*) vs. +SGQA (*purple*) across eight encoders spanning CLIP, DINOv2, and DINOv3 [37]. Red line (right axis): scoring gap ratio.

4.2 Analysis and Discussion

Unless otherwise stated, all experiments in this section use the COCO 30-shot setting. As established in the diagnostic analysis (Sec. 3.3), scoring quality accounts for 64% of the 15.1 AP oracle gap, which is why the analysis below focuses on the scoring stage.

Alternative Scoring Strategies. Among the six alternatives evaluated, SGQA gives the largest gain (+**2.6** AP) and adds no hyperparameters to the scoring formula. The arithmetic mean reaches only +1.1 AP, since its linear iso-contours still let a single high score offset a low one (Fig. 3), while strategies targeting other pipeline components, such as Soft-NMS, show negligible or negative effects.

Effect of the Weighting Parameter α. We sweep α in the generalized fusion (Eq. (3)) over $\{0.1, \ldots, 0.9\}$. Performance peaks at $\alpha = 0.5$ and falls off toward either extreme (Fig. 5), and this behavior is consistent across encoder and shot-count configurations, so we fix $\alpha = 0.5$ as the default and treat it as a constant rather than a tunable hyperparameter. A finer sweep with a step size of 0.05 retains the same optimum.

Different Encoder Architectures. SGQA improves results on all eight encoders tested, spanning CLIP [33], DINOv2 [28], and DINOv3 [37] (Fig. 6). *Quality Misalignment* therefore appears to be a recurring property of frozen compositions rather than an artifact of a particular encoder.

Additional Ablations. Across four SAM2 variants, SAM2-Tiny with SGQA (35.2 AP) approaches SAM2-Large without SGQA (36.8 AP), and the gain is similar across object scales (AP_S: +3.0, AP_M: +3.2, AP_L: +3.4).

Table 1: Comparison with state-of-the-art on COCO-FSOD (10/30-shot). nAP denotes mAP over novel classes. FT = finetuning on novel classes. Competing results from [8]. Best in **bold**, second best underlined. †DINOv2 ViT-L + SAM2 Hiera-L; others use ResNet-50/101 (25–45M) unless noted.

Method	FT	10-shot			30-shot		
		nAP	nAP_{50}	nAP_{75}	nAP	nAP_{50}	nAP_{75}
TFA [41]	✓	10.0	19.2	9.2	13.5	24.9	13.2
FSCE [39]	✓	11.9	–	10.5	16.4	–	16.2
Retentive RCNN [6]	✓	10.5	19.5	9.3	13.8	22.9	13.8
HeteroGraph [11]	✓	11.6	23.9	9.8	16.5	31.9	15.5
Meta F. R-CNN [12]	✓	12.7	25.7	10.8	16.6	31.8	15.8
LVC [18]	✓	19.0	34.1	19.0	26.8	45.8	27.5
C. Transformer [13]	✓	17.1	30.2	17.0	21.4	35.5	22.1
NIFF [9]	✓	18.8	–	–	20.9	–	–
DiGeo [27]	✓	10.3	18.7	9.9	14.2	26.2	14.8
CD-ViTO (ViT-L) [8]	✓	35.3	54.9	37.2	35.9	54.5	38.0
FSRW [17]	×	5.6	12.3	4.6	9.1	19.0	7.6
Meta R-CNN [46]	×	6.1	19.1	6.6	9.9	25.3	10.8
DE-ViT (L) [51]	×	34.0	53.0	37.0	34.0	52.9	37.2
NTT† [3]	×	36.6	54.1	38.3	36.8	54.5	38.7
SGQA (ours)†	×	**39.3**	**57.0**	**41.5**	**39.4**	**57.3**	**41.7**

4.3 Comparison with State-of-the-Art Methods

Table 1 reports COCO-FSOD results (10/30-shot, Bbox nAP/nAP50/nAP75 on novel classes). Without SGQA, NTT [3] is already competitive without any fine-tuning. Adding SGQA raises 30-shot nAP to 39.4, exceeding the best fine-tuning method CD-ViTO [8] by 3.5 nAP and the strongest prior published non-fine-tuning detector DE-ViT [51] by 5.4 nAP. The 10-shot and 30-shot results are close (39.3 vs. 39.4), and NTT shows the same pattern (36.6 vs. 36.8), which suggests that the K-Means prototypes already capture much of the intra-class variation with as few as 10 references; we read this as a property of prototype-based matching rather than an effect of SGQA. SGQA and NTT use larger frozen backbones (DINOv2 ViT-L + SAM2 Hiera-L, ∼524M parameters) than most competitors (ResNet-50/101), so the controlled comparison is NTT vs. SGQA, which share identical backbones and isolate the +2.6 nAP effect of quality alignment.

PASCAL VOC Few-Shot Benchmark. PASCAL VOC [4] comprises 20 classes (15 base, 5 novel) across three standard splits [9]; we report novel-class AP50 following [51]. As shown in Tab. 2, SGQA improves over the baseline on every split and shot setting, by +4.3 to +7.3 AP50 (avg. +6.2), and exceeds the prior fine-tuning and non-fine-tuning methods listed, which indicates that the effect carries over from COCO to VOC. The VOC gain (+6.2 AP50) is larger than the COCO gain (+2.6 nAP), which we attribute to the evaluation met-

Table 2: AP50 results on the novel classes of the PASCAL VOC few-shot benchmark. Competing method results are sourced from [51]. Best in **bold**, second best underlined. *The corresponding implementation is not publicly accessible. FT refers to fine-tuning on novel classes.

Method	FT	Novel Split 1					Novel Split 2					Novel Split 3					Avg
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
FsDetView [45]	✓	25.4	20.4	37.4	36.1	42.3	22.9	21.7	22.6	25.6	29.2	32.4	19.0	29.8	33.2	39.8	29.2
TFA [41]	✓	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8	39.9
Retentive RCNN [6]	✓	42.4	45.8	45.9	53.7	56.1	21.7	27.8	35.2	37.0	40.3	30.2	37.6	43.0	49.7	50.1	41.1
DiGeo [27]	✓	37.9	39.4	48.5	58.6	61.5	26.6	28.9	41.9	42.1	49.1	30.4	40.1	46.9	52.7	54.7	44.0
HeteroGraph [11]	✓	42.4	51.9	55.7	62.6	63.4	25.9	37.8	46.6	48.9	51.1	35.2	42.9	47.8	54.8	53.5	48.0
Meta Faster R-CNN [12]	✓	43.0	54.5	60.6	66.1	65.4	27.7	35.5	46.1	47.8	51.4	40.6	46.4	53.4	59.9	58.6	50.5
CrossTransformer [13]	✓	49.9	57.1	57.9	63.2	67.1	27.6	34.5	43.7	49.2	51.2	39.5	54.7	52.3	57.0	58.7	50.9
LVC [18]	✓	54.5	53.2	58.8	63.2	65.7	32.8	29.2	50.7	49.8	50.6	48.4	52.7	55.0	59.6	59.6	52.3
NIFF* [9]	✓	62.8	67.2	68.0	70.3	68.8	38.4	42.9	54.0	56.4	54.0	56.4	62.1	61.2	64.1	63.9	59.4
Multi-Relation Det [5]	×	37.8	43.6	51.6	56.5	58.6	22.5	30.6	40.7	43.1	47.6	31.0	37.9	43.7	51.3	49.8	43.1
DE-ViT (ViT-S/14) [51]	×	47.5	64.5	57.0	68.5	67.3	43.1	34.1	49.7	56.7	60.8	52.5	62.1	60.7	61.4	64.5	56.7
DE-ViT (ViT-B/14) [51]	×	56.9	61.8	68.0	73.9	72.8	45.3	47.3	58.2	59.8	60.6	58.6	62.3	62.7	64.6	67.8	61.4
DE-ViT (ViT-L/14) [51]	×	55.4	56.1	68.1	70.9	71.9	43.0	39.3	58.1	61.6	63.1	58.2	64.0	61.3	64.2	67.3	60.2
NTT [3]	×	70.8	72.3	73.3	77.2	79.1	54.5	67.0	76.3	75.9	78.2	61.1	67.9	71.3	70.8	72.6	71.2
SGQA (ours)	×	**76.5**	**79.2**	**79.9**	**82.5**	**83.4**	**60.6**	**73.3**	**82.1**	**82.3**	**84.1**	**67.5**	**74.2**	**78.3**	**77.1**	**79.9**	**77.4**

ric: AP50 uses a single IoU threshold of 0.5 and is more sensitive to score-level changes, whereas nAP averages over 0.5:0.05:0.95 and weights localization accuracy more heavily, diluting the benefit of scoring corrections. Matcher [26] and PerSAM [49] adopt different protocols (COCO-20^i mIoU and PerSeg), precluding direct nAP comparison without reimplementation; Sec. 3.3 instead offers a structural comparison via score fusion weights.

Multi-Seed Robustness. Few-shot methods are sensitive to which reference images are sampled, so we evaluate SGQA across 5 independent samplings of the reference set. SGQA yields a stable improvement of $+2.6 \pm 0.1$ AP over the NTT baseline, and its run-to-run variance stays close to the baseline (± 0.4 vs. ± 0.3 AP for NTT). SGQA is a deterministic post-hoc score correction and does not alter the reference sampling process or prototype construction, so it does not add to the variance already present in few-shot reference selection.

Low-Shot Scaling. Table 3 reports detection and segmentation metrics across shot counts from 1 to 30. The SGQA gain grows as the shot count decreases: +3.3 Bbox AP and +3.8 Segm AP at 1-shot, compared to +2.6 and +3.3 AP at 30-shot. This is consistent with our diagnosis: with fewer reference images, the K-Means prototypes are noisier estimates of the intra-class distribution, which makes the semantic score s^{sem} less reliable and the geometric-mean correction more useful, since it down-weights candidates whose high s^{sem} is not supported by geometric evidence. At the 1-shot setting, SGQA reaches 36.8 Bbox AP, matching the NTT 30-shot baseline.

Power Mean Comparison. To place the geometric mean within the broader aggregation family, we evaluate $M_p = \left(\frac{(s^{\text{sem}})^p + (s^{\text{geo}})^p}{2}\right)^{1/p}$. We consider $p \in \{-2, -1, -0.5, 0, 0.5, 1, 2\}$, where $p \to 0$ recovers the geometric mean. The geometric mean (p=0) achieves the highest AP (39.4), while the arithmetic mean

Table 3: Low-shot scaling on COCO few-shot (novel classes). SGQA improvement increases as shots decrease.

Shots	Bbox nAP		Δ_{box}	Segm nAP		Δ_{seg}
	NTT	SGQA		NTT	SGQA	
1	33.5	36.8	**+3.3**	30.8	34.6	**+3.8**
5	35.8	38.8	+3.0	33.2	36.7	+3.5
10	36.6	39.3	+2.7	34.0	37.3	+3.3
30	36.8	**39.4**	+2.6	34.2	**37.5**	+3.3

(p=1) yields only +1.1 AP because its linear iso-contours let one high score compensate for a low one, the misalignment pattern identified in Sec. 3.2. Stronger penalties (p<0, e.g., the harmonic mean at $p = -1$) suppress valid detections whose geometric quality is slightly below their semantic relevance, reducing the net gain. The geometric mean therefore penalizes score imbalance without over-suppressing geometrically imprecise but semantically correct candidates.

Factorial Analysis. SGQA uses the unified quality score Q in two pipeline roles at once: the detection confidence reported to the evaluator, and the ranking criterion used during NMS. To separate these contributions, we conduct a 2×2 factorial experiment that toggles each substitution independently. Using Q only for the NMS ranking score yields +1.7 AP, while using it only for the confidence yields +1.3 AP, so reranking during suppression is the larger of the two channels. The sum of the individual effects (+3.0 AP) exceeds the joint effect (+2.6 AP), indicating mild sub-additivity: a better NMS ranking already promotes stronger candidates to the top, which overlaps with the benefit of a more calibrated confidence.

Upper-Bound Analysis. The score-level-recoverable gap, defined as the largest improvement attainable by any monotone function of s^{sem} and s^{geo}, is estimated at 6.2 AP. SGQA recovers 42% of this gap (2.6 of 6.2 AP), corresponding to 17% of the full oracle gap (2.6 of 15.1 AP). Closing the remaining 58% of the score-level gap would require feature-level information exchange between the encoder and the generator, which we leave to future work.

Qualitative Results. Figure 7 compares NTT and SGQA on COCO 30-shot, showing three patterns that follow from the geometric mean's penalty on score imbalance (Sec. 3.4). First, candidates with high s^{sem} but poor geometric fidelity, the off-diagonal cases in Fig. 3, are demoted, which removes spurious detections. Second, by rewarding both quality dimensions jointly, SGQA favors masks with tighter instance boundaries over spatially imprecise alternatives. Third, partially occluded objects with moderate scores on both axes are ranked above unoccluded false positives that score high on only one axis. These cases match the dominant failure mode described in Sec. 3.3.

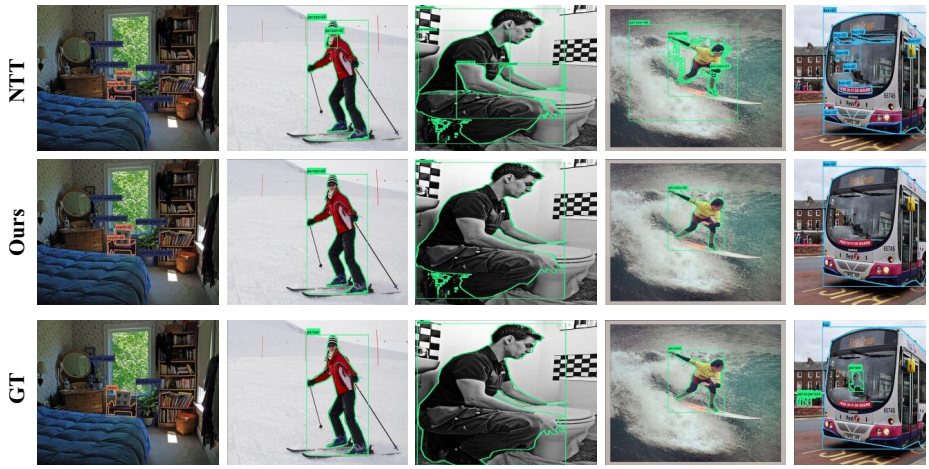

Fig. 7: Qualitative comparison on COCO 30-shot. Top row: NTT baseline; middle row: SGQA (ours); bottom row: ground truth. Compared with NTT, SGQA produces fewer false positives and tighter mask boundaries, including on occluded objects.

5 Conclusion

We study why composing frozen foundation models underperforms on training-free few-shot instance segmentation. Using progressive oracle replacement together with a comparative analysis of the scoring stage across several encoders and mask generators, we locate the largest loss at the scoring stage, which accounts for 64% of the performance gap. We attribute this to *Quality Misalignment*, a recurring mismatch between semantic matching confidence and mask geometric fidelity that we observe across all tested configurations and that parallels the classification-localization misalignment in supervised detection. Following this analysis, we introduce SGQA, a hyperparameter-free geometric-mean fusion of the two scores. SGQA recovers 42% of the score-level gap (2.6 of 6.2 AP) and 17% of the full oracle gap (2.6 of 15.1 AP). It surpasses existing training-free methods and outperforms fine-tuning-based methods with comparable or smaller backbones, indicating that aligning the two quality signals at the scoring stage suffices to improve frozen compositions without retraining.

Acknowledgements

This work was supported by the Advanced Materials-National Science and Technology Major Project (Grant No. 2025ZD0620100), and the AI for Science Program, Shanghai Municipal Commission of Economy and Informatization (Grant No. 2025-GZL-RGZN-BTBX-02033).

References

1. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms–improving object detection with one line of code. In: Proceedings of the IEEE international conference on computer vision. pp. 5561–5569 (2017)
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
3. Espinosa, M., Yang, C., Ericsson, L., McDonagh, S., Crowley, E.J.: No time to train! training-free reference-based instance segmentation. arXiv preprint arXiv:2507.02798 (2025)
4. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
5. Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W.: Few-shot object detection with attention-rpn and multi-relation detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4013–4022 (2020)
6. Fan, Z., Ma, Y., Li, Z., Sun, J.: Generalized few-shot object detection without forgetting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4527–4536 (2021)
7. Feng, C., Zhong, Y., Gao, Y., Scott, M.R., Huang, W.: Tood: Task-aligned one-stage object detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3490–3499. IEEE Computer Society (2021)
8. Fu, Y., Wang, Y., Pan, Y., Huai, L., Qiu, X., Shangguan, Z., Liu, T., Fu, Y., Van Gool, L., Jiang, X.: Cross-domain few-shot object detection via enhanced open-set object detector. In: European Conference on Computer Vision. pp. 247–264. Springer (2024)
9. Guirguis, K., Meier, J., Eskandar, G., Kayser, M., Yang, B., Beyerer, J.: Niff: Alleviating forgetting in generalized few-shot object detection via neural instance feature forging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24193–24202 (2023)
10. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
11. Han, G., He, Y., Huang, S., Ma, J., Chang, S.F.: Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3263–3272 (2021)
12. Han, G., Huang, S., Ma, J., He, Y., Chang, S.F.: Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 780–789 (2022)
13. Han, G., Ma, J., Huang, S., Chen, L., Chang, S.F.: Few-shot object detection with fully cross-transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5321–5330 (2022)
14. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: European conference on computer vision. pp. 340–353. Springer (2012)
15. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727. Springer (2022)
16. Jiang, B., Luo, R., Mao, J., Xiao, T., Jiang, Y.: Acquisition of localization confidence for accurate object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 784–799 (2018)

17. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8420–8429 (2019)
18. Kaul, P., Xie, W., Zisserman, A.: Label, verify, correct: A simple few shot object detection method. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14237–14247 (2022)
19. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1-2), 81–93 (1938)
20. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
21. Lan, M., Chen, C., Ke, Y., Wang, X., Feng, L., Zhang, W.: Clearclip: Decomposing clip representations for dense vision-language inference. In: European Conference on Computer Vision. pp. 143–160. Springer (2024)
22. Li, K., Liu, R., Cao, X., Bai, X., Zhou, F., Meng, D., Wang, Z.: Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10545–10556 (2025)
23. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. Advances in neural information processing systems **33**, 21002–21012 (2020)
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
25. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
26. Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment anything with one shot using all-purpose feature matching. In: International Conference on Learning Representations. vol. 2024, pp. 17904–17925 (2024)
27. Ma, J., Niu, Y., Xu, J., Huang, S., Han, G., Chang, S.F.: Digeo: Discriminative geometry-aware learning for generalized few-shot object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3208–3218 (2023)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features Without Supervision. Transactions on Machine Learning Research (2024), https://mlanthology.org/tmlr/2024/oquab2024tmlr-dinov2/
29. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)
30. Qiao, L., Zhao, Y., Li, Z., Qiu, X., Wu, J., Zhang, C.: Defrcn: Decoupled faster r-cnn for few-shot object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8681–8690 (2021)

31. Qing, Y., Wang, Y., Chen, C., Zhang, W., Wen, J., Xu, X.: S2c2seg: Semantic-spatial consistency and category optimization for open-vocabulary segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6293–6303 (June 2026)
32. Qing, Y., Zeng, D., Xie, S., Huang, K., Wang, Y.: Integrating low-level visual cues for enhanced unsupervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6603–6611 (2025)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
34. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollar, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations. vol. 2025, pp. 28085–28128 (2025)
35. Ren, T., Chen, Y., Jiang, Q., Zeng, Z., Xiong, Y., Liu, W., Ma, Z., Shen, J., Gao, Y., Jiang, X., et al.: Dino-x: A unified vision model for open-world object detection and understanding. arXiv preprint arXiv:2411.14347 (2024)
36. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
37. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
38. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing **107**, 104117 (2021)
39. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: Fsce: Few-shot object detection via contrastive proposal encoding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7352–7362 (2021)
40. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
41. Wang, X., Huang, T., Gonzalez, J., Darrell, T., Yu, F.: Frustratingly simple few-shot object detection. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 9919–9928. PMLR (13–18 Jul 2020)
42. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6830–6839 (2023)
43. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Towards segmenting everything in context. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1130–1140 (2023)
44. Wang, Y., Zou, X., Yan, L., Zhong, S., Zhou, J.: Snida: Unlocking few-shot object detection with non-linear semantic decoupling augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12544–12553 (2024)
45. Xiao, Y., Lepetit, V., Marlet, R.: Few-shot object detection and viewpoint estimation for objects in the wild. IEEE Transactions on Pattern Analysis and Machine

Intelligence **45**(3), 3090–3106 (2023). https://doi.org/10.1109/TPAMI.2022.3174072
46. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta r-cnn: Towards general solver for instance-level low-shot learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9577–9586 (2019)
47. Zhang, G., Luo, Z., Cui, K., Lu, S., Xing, E.P.: Meta-detr: Image-level few-shot detection with inter-class correlation exploitation. IEEE transactions on pattern analysis and machine intelligence **45**(11), 12832–12843 (2022)
48. Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N.: Varifocalnet: An iou-aware dense object detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8514–8523 (2021)
49. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Qiao, Y., Peng, G., Li, H.: Personalize segment anything model with one shot. In: International Conference on Learning Representations. vol. 2024, pp. 18250–18279 (2024)
50. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9759–9768 (2020)
51. Zhang, X., Liu, Y., Wang, Y., Boularias, A.: Detect everything with few examples. In: Proceedings of The 8th Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 270, pp. 3986–4004. PMLR (06–09 Nov 2025)