BOUNDARY-INFORMED SUPPORT-GUIDED EVIDENCE FUSION FOR TRAINING-FREE MULTIMODAL ONE-SHOT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

How should category text and visual support be combined when they disagree about the identity or extent of a query object? We introduce Support-Guided Evidence Fusion (SGEF), which treats the annotated support as both a prompt and a reliability reference. A frozen SAM3 evaluates one query-text view and two layouts of the same one-shot support–query pair. Their aligned fields form a semantic-rescue branch and a layout-consistent signed-correction branch, while support-region overlap and inner–outer boundary fidelity determine an episode-specific mixing weight. This boundary-informed rule adds no learned parameters and requires no task-specific training or test-time optimization. Across eight benchmarks, SGEF reaches 89.28, 77.42, and 61.82 foreground mIoU on PASCAL, COCO, and LVIS, and obtains the highest score among the evaluated methods on four of five additional benchmarks. Component replacements and episode-level weight reassignment isolate the contributions of directional correction and support-boundary fidelity. Across six frozen encoder–decoder configurations evaluated on complete PASCAL and COCO fold 0, fusion improves the corresponding reference field in all twelve comparisons by 0.57–1.85 mIoU points. These results establish annotated support boundaries as an effective control signal for training-free multimodal one-shot segmentation.

1 INTRODUCTION

One-shot segmentation uses an annotated image to specify a target in a distinct query image (Shaban et al., 2017). Visual support describes its appearance and extent, while category text supplies semantic knowledge acquired during language–image pretraining (Radford et al., 2021). These cues can disagree when support appearance changes across scenes or text identifies a distracting instance. Even when both cues identify the target, they may disagree over its extent. Adding foreground can recover an omission or introduce leakage; removing it can suppress leakage or erase the target. A fusion rule must therefore determine where to add or remove evidence and how strongly to apply that change.

Frozen promptable models produce complementary predictions from text and geometric prompts. SAM3 (Carion et al., 2026) represents both as foreground probability fields. Presenting support and query on a shared canvas exposes the annotated exemplar and the unlabelled query to the same frozen model; changing their layout supplies complementary support-conditioned estimates alongside the query-text prediction. These views can agree while making the same error. The query has no annotation at inference, leaving the strength of a proposed correction unresolved by agreement alone.

The annotated support supplies a spatial reference for this decision. Paired inference predicts both support and query, allowing the support prediction to be checked against a known mask within the same view. Region overlap measures coverage across the object. Inner and outer contour bands assess foreground retention and background rejection locally, where area overlap can dilute the effect of errors. Together, these measurements determine a fusion weight for the episode.

Support-Guided Evidence Fusion (SGEF) couples this weight to two complementary constructions from the aligned query predictions. Semantic rescue preserves the paired estimate and adds stronger query-text evidence. The signed branch uses residuals that permit additions or removals when the

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.

two paired layouts agree on the direction. The query fields define the spatial difference between these branches; support fidelity controls its contribution to the final prediction.

SGEF implements this division through fixed operations on frozen model outputs. The evaluation connects eight-benchmark segmentation performance with component replacements that test the correction rule and the support measurements. Weight reassignment examines whether reliability should remain associated with its episode, and six alternative frozen configurations test the fusion construction beyond SAM3.

2 Related Work

Support-conditioned segmentation. Few-shot segmentation transfers sparse annotations through prototypes, enriched priors, correlations, and transport-based matching (Wang et al., 2019; Tian et al., 2022; Luo et al., 2024; Min et al., 2021; Peng et al., 2023; Liu et al., 2023; Cheng et al., 2023; Lang et al., 2023; Bi et al., 2025). Cross-domain extensions transform features or mine support–query correspondence (Lei et al., 2022; Huang et al., 2023; Nie et al., 2024). Recent training-free methods transfer visual support through matching, prompting, or frozen features (Liu et al., 2024; Zhang et al., 2024a;b; Zakir & Ho, 2026; Lee et al., 2026; Cuttano et al., 2026; Herzog, 2024). SGEF uses category text alongside visual support and evaluates the resulting predictions against the support mask to regulate fusion.

Multimodal and spatial prompting. SegGPT formulates segmentation as in-context coloring, while the Segment Anything family develops promptable image, video, and concept segmentation (Wang et al., 2023; Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026). FSS-SAM3 uses a paired canvas to present an annotated exemplar alongside the query, with text providing a semantic target description (Tsai et al., 2026). SGEF uses two such layouts alongside a query-text view. Their predictions supply both the query fields to be fused and the support fields used to assess reliability.

Evidence fusion for segmentation. Support-conditioned weighting also appears in MM-FSS (An et al., 2024) for few-shot 3D segmentation, where TACC uses support-prediction IoU to modulate guidance. CDAM (Kang et al., 2025) builds attention from similarities between patch-wise class distributions for open-vocabulary segmentation. DecAF (Han et al., 2025) combines object/background and video/frame attention for training-free video-reasoning segmentation. SGEF evaluates region overlap and inner/outer contour fidelity against the annotated support. The resulting scalar controls a mixture of semantic rescue and layout-consistent signed correction in the aligned query fields.

3 Method

3.1 Overview and task formulation

An episode contains a support image I_s, its complete target mask M_s, a distinct query image I_q, and category text t. The output is the query mask $\widehat{M}_q$ in the query's image coordinates. The tight box enclosing M_s prompts SAM3, while the complete mask remains the support-fidelity reference. All SAM3 parameters are frozen. Query annotations are used after inference to compute evaluation metrics.

Figure 1 traces three independently evaluated views sharing frozen parameters: query-text, vertical support–query, and horizontal support–query. The query outputs form rescue R and signed correction C. The paired support outputs determine a scalar weight w, yielding the final field $E = (1 - w)R + wC$.

3.2 Aligned multimodal evidence

For object j, let $p_j(x) \in [0, 1]$ denote its native probability field and $s_j \in [0, 1]$ its confidence. A paired view uses category text and the support box; the query-text view has no geometric prompt. We crop each canvas region, take the score-weighted maximum $\max_j s_j p_j$, and bilinearly remap

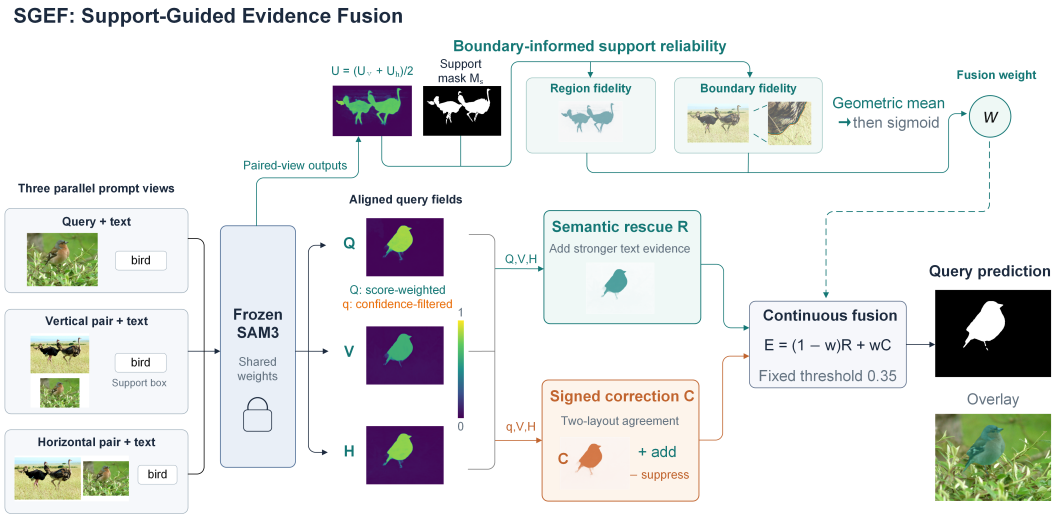

Figure 1: SGEF overview. A frozen SAM3 evaluates a query–text view and two paired support–query layouts. Their aligned fields form semantic rescue R and signed correction C. Region and boundary fidelity on the annotated support determine the mixing weight w (dashed connector), and the fused field E is thresholded at 0.35. Both Q and q come from the same query–text inference through different reductions. Images and fields come from one PASCAL episode; field colors use a linear 0–1 scale.

the result to the corresponding image coordinates. These operations yield query fields Q, V, H and support fields U_v, U_h. An empty object set yields a zero field.

The paired mean is $P = (V + H)/2$. The query-text output also yields q for signed correction. We retain object fields with $s_j > 0.40$, take their unweighted maximum, and apply the same remapping. If no object passes this fixed confidence cutoff, $q = 0$. Thus, Q is score-weighted, whereas q is confidence-filtered; both are derived from the same query-text inference. Below, $[z]_+ = \max(z, 0)$, $\mathrm{clip}_{[0,1]}$ clips to $[0, 1]$, and products between fields are pointwise.

3.3 Semantic Rescue

Semantic rescue supplements the paired estimate with stronger query-text evidence. We first measure their fuzzy overlap,

$$a = \frac{\sum_x \min(Q(x), P(x))}{\sum_x \max(Q(x), P(x))}. \tag{1}$$

We set $a = 1$ when the denominator is at most 10^{-6} and use it to scale positive text residuals,

$$R = \mathrm{clip}_{[0,1]}\left(P + (1 - a)^3 [Q - P]_+\right). \tag{2}$$

Since $P \le R \le \max(P, Q)$ pointwise, rescue preserves the paired field wherever text evidence is weaker. Higher global overlap attenuates additions, while the signed branch permits reductions relative to its query-text input.

3.4 Layout-Consistent Signed Correction

To allow reductions as well as additions, we compare the confidence-filtered text field q with both paired fields,

$$d_+ = \sqrt{[V - q]_+ [H - q]_+}, \qquad d_- = \sqrt{[q - V]_+ [q - H]_+}. \tag{3}$$

An addition requires both V and H to exceed q; a reduction requires both to fall below it. If q lies between the paired fields, both residuals vanish. Exchanging the two layouts leaves the residuals

unchanged, and $d_+ d_- = 0$ prevents both residual directions from being active at a pixel. The signed branch is then

$$\alpha = 0.5 + 0.5 \exp(-\operatorname{mean}(q)/0.04), \tag{4}$$

$$S = \operatorname{clip}_{[0,1]} \left(q + \alpha d_+ - 0.75\sqrt{q}\, d_- \right), \tag{5}$$

$$C = \operatorname{clip}_{[0,1]}(S + 0.05). \tag{6}$$

Sparse text coverage increases the positive coefficient, while $\sqrt{q}$ attenuates removal at weak-evidence pixels. Agreement controls the signed residuals; the fixed offset raises the clipped field before a final clipping to $[0, 1]$. All coefficients are fixed across datasets; Appendix A specifies the corresponding control operations.

3.5 Support-region and boundary fidelity

Support fidelity measures agreement between a paired support prediction and the annotated target. We assess its overall extent and its behavior immediately inside and outside the contour. For the mean support field $U = (U_v + U_h)/2$, region fidelity is fuzzy intersection over union,

$$r_{\text{reg}} = \frac{\sum_x \min(U(x), M_s(x))}{\max(\sum_x \max(U(x), M_s(x)), 10^{-6})}. \tag{7}$$

To distinguish undercoverage from leakage near the support contour, we additionally construct inner and outer bands,

$$B_{\text{in}} = M_s \setminus \operatorname{erode}(M_s), \qquad B_{\text{out}} = \operatorname{dilate}(M_s) \setminus M_s. \tag{8}$$

Erosion and dilation use a disk of radius $\operatorname{clip}(\operatorname{round}(0.006\sqrt{\max(|M_s|, 1)}), 1, 5)$ pixels. Here, $|M_s|$ is the foreground area in support-field coordinates. Boundary fidelity measures foreground just inside the contour and background just outside it,

$$r_{\text{bd}} = \sqrt{\operatorname{mean}_{B_{\text{in}}} U \;\; \operatorname{mean}_{B_{\text{out}}} (1 - U)}. \tag{9}$$

Empty bands receive zero fidelity. We combine region and boundary measurements as $r = \sqrt{r_{\text{reg}} r_{\text{bd}}}$ and map this scalar to

$$w = \sigma((r - 0.55)/0.08). \tag{10}$$

Here, $\sigma(z) = 1/(1 + \exp(-z))$ is the sigmoid function. Higher fidelity gives C greater weight, allowing the layout-agreed residuals to exert more influence on the final field. Lower fidelity favors R, which preserves the paired estimate while admitting stronger query-text evidence. The same scalar w controls interpolation at every query pixel.

3.6 Output and implementation

The final prediction follows from

$$E = (1 - w)R + wC, \qquad \widehat{M}_q(x) = \mathbf{1}[E(x) \geq 0.35]. \tag{11}$$

Since $R, C \in [0, 1]$ and $0 < w < 1$, each value of E lies between its two branch values. The soft-field update relative to rescue is $E - R = w(C - R)$; changing w scales this difference without reversing its sign. Fusion operates on the three views' output fields; support-fidelity computation reuses both paired predictions without additional backbone evaluations. Paired layouts are constructed on SAM3's native 1008×1008 input canvas with a support split ratio of 0.6. Probability fields use bilinear remapping, whereas masks use nearest-neighbor remapping. Dataset adapters retain their prescribed image preparation and canvas order, with shared fusion coefficients and a final threshold of 0.35. Appendix A specifies these settings and the exact substitutions used in the controls.

4 Experiments

The experiments assess segmentation accuracy, the roles of layout-consistent signed correction and support fidelity, and the effect of changing frozen evidence providers. Matched-view controls examine the fusion operations, and timing measurements quantify their system cost.

4.1 EVALUATION SETUP

Datasets and episodes. We evaluate eight datasets with one annotated support image per episode. The class-partitioned benchmarks are PASCAL VOC (Everingham et al., 2010), COCO (Lin et al., 2014), and LVIS (Gupta et al., 2019). Additional evaluation uses FSS-1000 (Li et al., 2020), PerSeg (Zhang et al., 2024b), DAVIS (Pont-Tuset et al., 2017), ISIC (Codella et al., 2018), and Kvasir-SEG (Jha et al., 2020). PASCAL-5^i and COCO-20^i each use four folds of 1,000 episodes. LVIS-92^i uses ten folds of 2,300 episodes, with 25 episodes for each of 92 classes per fold. The additional query partitions contain 2,400 FSS-1000, 176 PerSeg, 1,969 DAVIS, 650 ISIC, and 200 Kvasir-SEG episodes.

Support and query images are distinct within each episode. PASCAL and COCO use complete target-class masks with deterministic episode construction. FSS-1000 uses deterministic same-class support sampling.

Metric aggregation. We map predictions to query coordinates and accumulate foreground intersection and union within each class. For fold f with class set $\mathcal{C}_f$, the score is

$$\mathrm{mIoU}_f = \frac{100}{|\mathcal{C}_f|} \sum_{c \in \mathcal{C}_f} \frac{\sum_{e:y_e=c} |\widehat{M}_e \cap M_e|}{\sum_{e:y_e=c} |\widehat{M}_e \cup M_e|}. \tag{12}$$

Here, y_e is the target class of episode e, and M_e is its query annotation. The class-level ratio pools intersections and unions before any averaging. Dataset means assign equal weight to each fold. Additional-benchmark evaluation pools within FSS-1000 categories, PerSeg object groups, and DAVIS sequences, then averages the group IoUs equally. ISIC and Kvasir-SEG each use a single foreground group, pooling intersections and unions across the complete query partition. Throughout the results, gains are expressed as absolute mIoU percentage-point differences. DAVIS reports support-conditioned frame foreground mIoU.

Comparison settings. All comparison results are from our evaluations of the listed methods. Each episode contains one annotated support image. SGEF renders that same one-shot support–query pair in three prompt views: one query-text view and two paired layouts. All three views reuse the episode's single annotated support image, preserving the one-shot protocol. FSS-SAM3 receives the same category text and annotated visual support through one paired view; measured inference costs are reported separately. CG-ICS (Chen et al., 2026) generates concept text from the support using Qwen3-VL-4B (Bai et al., 2025) and combines concept search with SAM3 geometric prompting. CG-ICS uses its native 512-pixel preprocessing and maps predictions to query coordinates on the same episode sequence. The remaining methods receive visual support without category text and retain their native inference or adaptation procedures. Equal-view fusion, symmetric correction, and region-only reliability retain SGEF's three input views; Table 7 reports system inference cost.

4.2 COMPARISON ON CLASS-PARTITIONED BENCHMARKS

SGEF achieved 89.28 mIoU on PASCAL and 77.42 on COCO, exceeding FSS-SAM3 by 2.87 and 3.60 points (Table 1). All four COCO folds improved over FSS-SAM3, with gains of 2.81–5.08 points. PASCAL gains ranged from 0.60 point on fold 1 to 8.33 points on fold 3. CG-ICS achieved 88.37 and 71.60 on PASCAL and COCO, respectively. SGEF exceeded these means by 0.91 and 5.82 points, although CG-ICS scored higher on PASCAL fold 3.

SGEF reached 61.82 mIoU on LVIS, compared with 53.13 for FSS-SAM3 (Table 2). Every fold improved, with gains between 5.96 and 9.89 points. Each of the 920 evaluated classes contributes 25 episodes, so the result covers all ten partitions under class-balanced sampling. CG-ICS achieved a ten-fold mean of 45.43, compared with 61.82 for SGEF, a difference of 16.39 points.

At class level, SGEF improved 80.5% of the 920 evaluated classes, with a median gain of 4.41 points. Classes with FSS-SAM3 IoU below 25 gained 17.89 points on average; those at or above 75 gained 1.49 points. Grouping classes by FSS-SAM3 accuracy therefore showed larger gains on its lower-scoring categories.

Table 1: One-shot foreground mIoU (%) on PASCAL-5^i and COCO-20^i: 1,000 episodes per fold. SGEF and FSS-SAM3 use category text and visual support; CG-ICS generates text from support. Other methods use visual support. Mean averages four folds; bold marks column maxima.

	PASCAL-5^i					COCO-20^i				
Method	F0	F1	F2	F3	Mean	F0	F1	F2	F3	Mean
PerSAM	50.18	53.35	45.27	39.98	47.20	22.62	23.74	21.40	20.79	22.14
PerSAM-F	46.20	51.03	46.69	38.31	45.56	21.85	23.59	23.13	22.34	22.73
FSSDINO	51.46	62.56	58.82	51.46	56.08	38.72	41.33	40.12	42.44	40.65
PR-MaGIC	66.65	67.36	63.17	65.54	65.68	50.46	53.65	56.32	53.13	53.39
INSID3	63.77	73.14	67.96	68.42	68.32	52.34	55.41	54.27	53.82	53.96
Matcher	67.66	70.29	66.05	67.69	67.92	51.48	52.76	51.22	52.72	52.05
GF-SAM	70.61	75.78	68.32	72.40	71.78	56.56	60.96	59.33	57.71	58.64
ABCDFSS	35.17	43.83	40.97	36.51	39.12	22.02	25.40	24.24	24.35	24.00
FSS-SAM3	86.72	92.59	86.71	79.64	86.41	73.59	73.33	72.56	75.80	73.82
CG-ICS	86.04	91.00	87.70	**88.73**	88.37	71.15	72.50	71.77	70.98	71.60
SGEF (ours)	**87.39**	**93.19**	**88.57**	87.97	**89.28**	**76.75**	**76.14**	**77.64**	**79.15**	**77.42**

Table 2: Foreground mIoU (%) on LVIS-92^i: 92 classes and 2,300 one-shot episodes per fold. SGEF and FSS-SAM3 use text and support; GF-SAM and Matcher use visual support. CG-ICS uses support-generated text. Mean averages all ten folds; bold marks column maxima.

Method	F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	Mean
FSS-SAM3	49.35	57.02	53.92	56.91	50.27	51.47	52.06	54.24	55.40	50.68	53.13
GF-SAM	30.64	37.98	37.20	39.54	33.27	35.13	35.74	37.86	37.90	35.23	36.05
Matcher	31.16	30.62	33.69	37.85	31.57	33.26	33.73	34.18	34.22	32.91	33.32
CG-ICS	43.94	49.84	45.37	49.33	42.08	44.68	43.28	44.12	45.37	46.27	45.43
SGEF (ours)	**59.24**	**64.30**	**59.88**	**66.20**	**59.12**	**61.19**	**60.94**	**62.23**	**64.94**	**60.18**	**61.82**

4.3 Evaluation on additional benchmarks

SGEF achieved 89.04 mIoU on FSS-1000, exceeding FSS-SAM3 by 4.49 points and CG-ICS by 0.63 point (Table 3). It reached 95.76 on PerSeg and 73.32 on DAVIS, with gains of 3.06 and 24.34 points over FSS-SAM3. On ISIC and Kvasir-SEG, SGEF obtained 65.34 and 76.83, exceeding CG-ICS by 2.54 and 0.42 points, respectively. SGEF had the highest mIoU among the evaluated methods on FSS-1000, PerSeg, ISIC, and Kvasir-SEG. SGEF and CG-ICS were nearly tied on PerSeg, differing by 0.01 point. On DAVIS, PR-MaGIC reached 79.58 and CG-ICS reached 75.63, exceeding SGEF by 6.26 and 2.31 points.

4.4 Component contributions

Table 4 compares SGEF with component replacements on FSS-1000 and PASCAL/COCO fold 0. Equal-view fusion, symmetric correction, and region-only reliability retain the complete model's three predictions and change their subsequent processing. The evidence-construction controls separate the available modalities. The query-text control used one query image with category text; the visual-support control used two geometry-only paired views without text. Visual support scored higher than query-text on all three evaluated partitions (Table 4). SGEF exceeded the visual-support scores by 2.10, 2.47, and 2.50 points on FSS-1000, PASCAL, and COCO, respectively. With all three multimodal views retained, equal-view fusion was lower than SGEF by 1.71, 1.53, and 1.98 points, respectively. This comparison uses the same evidence fields and attributes the measured difference to their subsequent processing.

Replacing the layout-consistent asymmetric update with symmetric correction reduced mIoU by 1.95, 1.21, and 1.43 points on FSS-1000, PASCAL, and COCO, respectively. The replacement retained both residual directions but assigned the fixed coefficient 0.75 to each. It removed the positive mass dependence and the $\sqrt{q}$ factor on removal, while preserving rescue, clipping, the offset, and the support weight.

Table 3: One-shot foreground mIoU (%) on five complete query partitions. SGEF and FSS-SAM3 use category text and support; CG-ICS generates text from support. Other methods use visual support. Bold marks column maxima. DAVIS measures support-conditioned frame segmentation.

Method	FSS-1000	PerSeg	DAVIS	ISIC	Kvasir-SEG
PerSAM	72.27	83.32	61.65	18.91	17.38
PerSAM-F	75.09	91.63	65.66	43.74	36.17
Matcher	85.12	86.85	72.61	50.69	35.96
GF-SAM	85.47	87.23	70.58	44.19	60.42
INSID3	83.07	90.14	73.53	52.16	43.92
FSSDINO	75.24	82.08	60.21	48.77	41.35
PR-MaGIC	84.69	94.13	**79.58**	53.97	56.82
ABCDFSS	69.16	82.84	56.37	44.18	55.09
FSS-SAM3	84.55	92.70	48.98	6.18	37.43
CG-ICS	88.41	95.75	75.63	62.80	76.41
SGEF (ours)	**89.04**	**95.76**	73.32	**65.34**	**76.83**

Table 4: Component ablations (foreground mIoU, %). FSS-1000 uses its full query partition; PASCAL and COCO use fold 0. Bold marks column maxima. Appendix A describes the control operations.

Setting	FSS-1000	PASCAL fold 0	COCO fold 0
Query image + text	86.23	83.15	73.14
Visual support (no text)	86.94	84.92	74.25
Equal multimodal views	87.33	85.86	74.77
Symmetric correction	87.09	86.18	75.32
Region-only reliability	87.64	85.71	76.13
SGEF (ours)	**89.04**	**87.39**	**76.75**

The support-reliability comparison isolates boundary information. Region-only reliability lowered mIoU by 1.40 on FSS-1000, 1.68 on PASCAL, and 0.62 on COCO relative to joint fidelity. The sigmoid and residual branches were unchanged; the substitution replaced $\sqrt{r_{\mathrm{reg}} r_{\mathrm{bd}}}$ with r_{reg}.

Finally, we tested whether reliability must remain associated with its episode. We compared the recorded weights with their fixed mean and five random permutations, holding R, C and the output threshold unchanged. Permutation preserved the weight distribution while changing its assignment to episodes. Table 5 reports the mean and range across the five permutations; Appendix A defines each substitution. Episode-matched weights exceeded the fixed mean by 1.86, 1.71, and 1.65 points on PASCAL, COCO, and LVIS, respectively. The corresponding gains over the permutation means were 1.90, 1.48, and 1.30 points. SGEF exceeded every tested permutation, including the permutation maxima by 0.47, 0.44, and 0.61 points.

4.5 Fusion with alternative frozen evidence providers

The transfer study combines OWL-ViT text proposals (Minderer et al., 2022) with DINOv2 or DINOv3 support-identity features (Oquab et al., 2023; Siméoni et al., 2026) and SAM or SAM2.1 mask decoders (Kirillov et al., 2023; Ravi et al., 2025). The six fixed configurations in Table 6 share the proposal rule, field construction, fusion constants, and output threshold. They use the listed identity encoders and mask decoders with their respective input preprocessing.

The reference field Q_A combines category-text scores with support-identity matches in each query image. The final field E_A, defined in Equation 17, applies support-controlled fusion to this field and the paired-view predictions. Both fields use the same frozen components and episodes. For each configuration, $\Delta = \mathrm{mIoU}(E_A) - \mathrm{mIoU}(Q_A)$ measures the resulting change; Appendix B defines the readout.

Table 5: Weight-assignment controls on fold 0 (foreground mIoU, %). Permutation entries report the mean across five random permutations, with the minimum and maximum in brackets. Bold marks the best score in each column.

Weight assignment	PASCAL fold 0	COCO fold 0	LVIS fold 0
Fixed mean	85.53	75.04	57.59
Random permutation	85.49 [84.06, 86.92]	75.27 [74.23, 76.31]	57.94 [57.24, 58.63]
SGEF (episode-matched)	**87.39**	**76.75**	**59.24**

Table 6: Alternative frozen evidence providers on complete fold 0 (1,000 episodes; foreground mIoU, %). Q_A is the support-identity query field and E_A the final output. $\Delta = \mathrm{mIoU}(E_A) - \mathrm{mIoU}(Q_A)$.

		PASCAL fold 0			COCO fold 0		
Encoder	Decoder	Q_A	E_A	Δ	Q_A	E_A	Δ
DINOv3-B	SAM2.1-Tiny	82.84	84.03	+1.19	67.29	68.77	+1.48
DINOv3-B	SAM2.1-Base+	82.51	84.35	+1.84	68.74	69.31	+0.57
DINOv3-B	SAM2.1-Large	84.72	85.96	+1.24	69.94	70.63	+0.69
DINOv3-B	SAM ViT-H	81.81	83.66	+1.85	68.30	70.06	+1.76
DINOv2-L	SAM2.1-Large	85.32	86.89	+1.57	71.50	73.04	+1.54
DINOv2-L	SAM ViT-H	82.28	83.95	+1.67	69.12	70.92	+1.80

The final readout improved the reference field in all twelve comparisons (Table 6). Gains ranged from 1.19 to 1.85 points on PASCAL and from 0.57 to 1.80 points on COCO. DINOv2-L with SAM2.1-Large achieved the highest final scores, reaching 86.89 and 73.04, respectively.

The highest alternative scores were 0.50 and 3.71 points below native SGEF on PASCAL fold 0 and COCO fold 0, respectively. Appendix B defines the frozen modules, field construction, and output equation.

4.6 Mechanism visualization

In Figure 2, rescue, region-only fusion, and SGEF achieved 53.54, 80.69, and 81.80 IoU; with branch fields fixed, joint fidelity lowered w from 0.967 to 0.801 and added 1.11 points over region-only fusion.

4.7 Deployment cost

Table 7 compares parameter count and serial query time on one RTX 3090. SGEF instantiated 840.51 million parameters and required 1.327 seconds per query: slower than single-view FSS-SAM3, but faster than PerSAM-F, PR-MaGIC, Matcher, and CG-ICS. FSSDINO was fastest, ABCDFSS instantiated the fewest parameters, and CG-ICS had the largest footprint. Appendix C gives the timing protocol and input perturbations; Figure 3 shows selected qualitative examples.

5 Discussion

SGEF treats the support annotation as both a prompt and a reliability reference. Query fields determine the correction's spatial form, whereas the known support mask controls its strength; by $E - R = w(C - R)$, lower fidelity contracts the output toward rescue without reversing the signed correction.

Region overlap and boundary fidelity constrain different support errors. The former measures object-level recovery; the latter tests both sides of the annotated contour. A thin boundary leak can change area overlap little while admitting background pixels immediately outside the object. The region-only control removes this contour test without changing the query branches and loses on FSS-1000, PASCAL, and COCO. Fixed-mean weighting removes episode variation; permutation misassigns

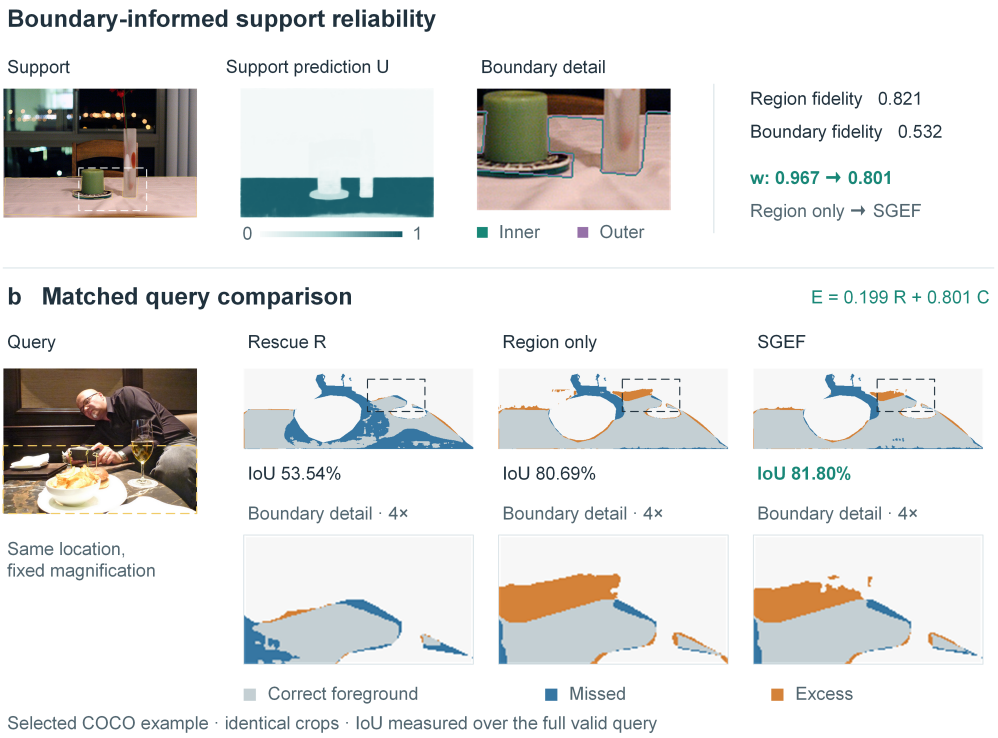

Figure 2: Boundary-informed reliability on a selected COCO episode. "Inner" and "Outer" are bands immediately inside and outside the support-mask contour, measuring foreground retention and background rejection. Joint region–boundary fidelity reduces w from 0.967 to 0.801. Below, gray, blue, and orange denote correct, missed, and excess foreground at threshold 0.35; dashed boxes show identical crops. Query annotations define the diagnostic error maps, and IoUs are computed over the full query.

measured weights. Both retain the branch fields and lower accuracy. Together with the symmetric-correction ablation, these interventions separate the direction, measurement, and episode assignment of a correction.

The complete-fold provider study probes a different aspect of generality. The same fixed readout improves all twelve encoder–decoder comparisons, although their best final scores remain below native SGEF. The consistent gains support the portability of the fusion rule, while the final-score differences show that absolute accuracy remains provider-dependent.

DAVIS marks a boundary of the observed gains: SGEF trails PR-MaGIC and CG-ICS on frame mIoU despite leading on four other additional benchmarks. The relative contributions of frame appearance change, support mismatch, and episode-level reliability on DAVIS remain unresolved. One scalar cannot express spatially varying trust, and three views increase inference cost. A natural extension is a spatial reliability field $w(x)$ that scales the signed correction locally. An adaptive policy could request another view only when needed; both extensions require reliability–outcome calibration and cost–accuracy evaluation.

6 Conclusion

SGEF combines layout-consistent signed correction with support-region and boundary fidelity for training-free multimodal one-shot segmentation, reaching 89.28, 77.42, and 61.82 foreground mIoU on PASCAL, COCO, and LVIS. Ablations, weight reassignment, and twelve complete-fold provider comparisons identify support boundaries as a control signal for reconciling text and visual evidence without additional learning.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the inference equations, fixed constants, preprocessing, metric aggregation, episode counts, and comparison conditions. Appendices A–C provide the control substitutions, frozen-provider definitions, timing boundary, and input-sensitivity settings used in the reported experiments.

AI USE STATEMENT

Generative AI tools were used for manuscript polishing and review. The authors take responsibility for the final manuscript.

REFERENCES

Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Min Wu, Ming-Ming Cheng, Ender Konukoglu, and Serge J. Belongie. Multimodality helps few-shot 3d point cloud semantic segmentation. *ArXiv*, abs/2410.22489, 2024. URL https://api.semanticscholar.org/CorpusID:273696085.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Hanbo Bi, Yingchao Feng, Wenhui Diao, Peijin Wang, Yongqiang Mao, Kun Fu, Hongqi Wang, and Xian Sun. Prompt-and-transfer: Dynamic class-aware enhancement for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):131–148, 2025. doi: 10.1109/TPAMI.2024.3461779.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Dídac Surís, Chaitanya K. Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Radle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e0982cbc81401df3430ee1ff780dc7a2-Abstract-Conference.html.

Zhigang Chen, Xiawu Zheng, and Rongrong Ji. Toward robust in-context segmentation via concept guidance. *ArXiv*, abs/2606.28149, 2026. URL https://api.semanticscholar.org/CorpusID:289669332.

Gong Cheng, Chunbo Lang, and Junwei Han. Holistic prototype activation for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4650–4666, 2023. doi: 10.1109/TPAMI.2022.3193587.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172. IEEE, 2018. doi: 10.1109/ISBI.2018.8363547. URL https://doi.org/10.1109/ISBI.2018.8363547.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. Insid3: Training-free in-context segmentation with dinov3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Cuttano_INSID3_Training-Free_In-Context_Segmentation_with_DINOv3_CVPR_2026_paper.pdf.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88 (2):303–338, June 2010. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-009-0275-4. URL http://link.springer.com/10.1007/s11263-009-0275-4.

Agrim Gupta, Piotr Dollár, and Ross B. Girshick. Lvis: A dataset for large vocabulary instance segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5351–5359, 2019. URL https://api.semanticscholar.org/CorpusID:195441339.

Su Ho Han, Jeongseok Hyun, Pilhyeon Lee, Minho Shim, Dongyoon Wee, and Seon Joo Kim. Decomposed attention fusion in mllms for training-free video reasoning segmentation. *ArXiv*, abs/2510.19592, 2025. URL https://api.semanticscholar.org/CorpusID:282272595.

Jonas Herzog. Adapt before comparison: A new perspective on cross-domain few-shot segmentation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23605–23615, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Herzog_Adapt_Before_Comparison_A_New_Perspective_on_Cross-Domain_Few-Shot_Segmentation_CVPR_2024_paper.html.

Xinyang Huang, Chuanglu Zhu, and Wenkai Chen. RestNet: Boosting Cross-Domain Few-Shot Segmentation with Residual Transformation Network. In *British Machine Vision Conference*, 2023. URL https://proceedings.bmvc2023.org/12/.

Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. *Kvasir-SEG: A Segmented Polyp Dataset*, pp. 451–462. Springer International Publishing, 2020. doi: 10.1007/978-3-030-37734-2_37. URL https://doi.org/10.1007/978-3-030-37734-2_37.

Dong-Un Kang, Hayeon Kim, and Se-Young Chun. Class distribution-induced attention map for open-vocabulary semantic segmentations. In *International Conference on Learning Representations*, 2025. URL https://api.semanticscholar.org/CorpusID:278532860.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment Anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023. doi: 10.1109/ICCV51070.2023.00371. URL https://doi.org/10.1109/ICCV51070.2023.00371.

Chunbo Lang, Gong Cheng, Binfei Tu, Chao Li, and Junwei Han. Base and meta: A new perspective on few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (9):10669–10686, 2023. doi: 10.1109/TPAMI.2023.3265865.

Minjae Lee, Sungwoo Hur, Soojin Hwang, and Won Hwa Kim. PR-MaGIC: Prompt refinement via mask decoder gradient flow for in-context segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21659–21668, 2026. URL https://arxiv.org/abs/2604.12113.

Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-Domain Few-Shot Semantic Segmentation. In *European Conference on Computer Vision*, 2022. doi: 10.1007/978-3-031-20056-4_5. URL https://doi.org/10.1007/978-3-031-20056-4_5.

Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. FSS-1000: A 1000-Class Dataset for Few-Shot Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2869–2878, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Li_FSS-1000_A_1000-Class_Dataset_for_Few-Shot_Segmentation_CVPR_2020_paper.html.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. URL https://api.semanticscholar.org/CorpusID:14113767.

Weide Liu, Chi Zhang, Henghui Ding, Tzu-Yi Hung, and Guosheng Lin. Few-Shot Segmentation With Optimal Transport Matching and Message Flow. *IEEE Transactions on Multimedia*, 25: 5130–5141, 2023. ISSN 1520-9210, 1941-0077. doi: 10.1109/TMM.2022.3187855.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment Anything with One Shot Using All-Purpose Feature Matching. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=yzRXdhk2he.

Xiaoliu Luo, Zhuotao Tian, Taiping Zhang, Bei Yu, Yuan Yan Tang, and Jiaya Jia. PFENet++: Boosting few-shot semantic segmentation with the noise-filtered context-aware prior mask. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1273–1289, 2024. doi: 10.1109/TPAMI.2023.3329725.

Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation Squeeze for Few-Shot Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6941–6952, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Min_Hypercorrelation_Squeeze_for_Few-Shot_Segmentation_ICCV_2021_paper.html.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers, 2022. URL https://arxiv.org/abs/2205.06230.

Jiahao Nie, Yun Xing, Gongjie Zhang, Pei Yan, Aoran Xiao, Yap-Peng Tan, Alex Chichung Kot, and Shijian Lu. Cross-domain few-shot segmentation via iterative support-query correspondence mining. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3380–3390, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *ArXiv*, abs/2304.07193, 2023. URL https://arxiv.org/abs/2304.07193.

Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical Dense Correlation Distillation for Few-Shot Segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23641–23651, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.02264.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS Challenge on Video Object Segmentation. *ArXiv*, abs/1704.00675, 2017. URL https://arxiv.org/abs/1704.00675.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ArXiv*, abs/2103.00020, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya K. Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao Wu, Ross B. Girshick, Piotr Dollár, and Christoph

Feichtenhofer. SAM 2: Segment Anything in Images and Videos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Amirreza Shaban, Shray Bansal, Z. Liu, Irfan Essa, and Byron Boots. One-Shot Learning for Semantic Segmentation. *ArXiv*, abs/1709.03410, 2017. URL https://arxiv.org/abs/1709.03410.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=2NlGyqNjns. Featured Certification.

Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior Guided Feature Enrichment Network for Few-Shot Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):1050–1065, February 2022. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2020.3013717. URL https://ieeexplore.ieee.org/document/9154595/.

Yi-Jen Tsai, Yen-Yu Lin, and Chien-Yao Wang. Few-shot semantic segmentation meets SAM3. *arXiv preprint arXiv:2604.05433*, 2026. URL https://arxiv.org/abs/2604.05433.

Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. PANet: Few-Shot Image Semantic Segmentation With Prototype Alignment. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9196–9205. IEEE, 2019. doi: 10.1109/ICCV.2019.00929. URL https://doi.org/10.1109/ICCV.2019.00929.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Towards Segmenting Everything in Context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1130–1140, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_SegGPT_Towards_Segmenting_Everything_in_Context_ICCV_2023_paper.html.

Hussni Mohd Zakir and Eric Tatt Wei Ho. Revealing the semantic selection gap in DINOv3 through training-free few-shot segmentation. *arXiv preprint arXiv:2602.07550*, 2026. URL https://arxiv.org/abs/2602.07550.

Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. In *Advances in Neural Information Processing Systems*, volume 37, pp. 33232–33261, 2024a. doi: 10.52202/079017-1047. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3a2ef31a1e45908901adc0ca853a8faf-Abstract-Conference.html.

Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Personalize Segment Anything Model with One Shot. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=6Gzkhoc6YS.