Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Training-free Boundary-First one-shot segmentation by closed-region hypothesis testing

EnBao Zhang

enbaozhang52@gmail.com

Hefei University of Economics https://orcid.org/0009-0008-2690-969X

Article

Keywords:

Posted Date: September 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10807399/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Training-free Boundary-First one-shot segmentation by closed-region hypothesis testing

EnBao Zhang*

Hefei University of Economics, Hefei 230012, Anhui, China

*Correspondence: enbaozhang52@gmail.com

Abstract

Most segmentation systems infer object extent from dense semantic decisions, leaving spatial coherence as an output rather than a decision constraint. We introduce Boundary-First segmentation, which constructs closed-region hypotheses before support-conditioned identity arbitration. Unified Boundary Evidence Transport (UBET) implements this ordering without training, using frozen region proposals, analytic DINOv2 identity evidence and a geometric conflict certificate. Across seven natural-image, object, video and medical benchmarks, SAM2- and visual-only SAM3-based UBET achieve leading or competitive one-shot segmentation. Six independent implementations using diffusion, depth, superpixel and conventional edge sources further demonstrate the portability of the paradigm. Decision-order interventions, candidate-ranking analyses and oracle diagnostics separate contour availability from identity arbitration. These results establish closed-region hypothesis testing as a transferable basis for future trainable or self-learning segmentation systems. Code is available in the public GitHub repository.

1 Introduction

Image segmentation is commonly formulated as dense labelling: a model assigns a class to each pixel, and an object mask emerges from these local decisions. This formulation has driven substantial progress in few-shot visual recognition, medical imaging and video understanding[1]. Yet object extent remains implicit in the pixel-wise prediction space. A locally plausible appearance match need not form a coherent object, especially when support and query images differ in scale, texture, illumination or background.

Foundation models provide useful ingredients for addressing this ambiguity. SAM and SAM2 generate class-agnostic region proposals, while DINOv2 supplies transferable visual correspondences[2–4]. Personalised and one-shot systems such as PerSAM and Matcher use these representations for prompt construction, feature matching and mask refinement[5,6]. These methods commonly use semantic evidence to initiate or constrain mask decoding. Object boundaries usually remain outputs, refinement cues or implicit consequences of dense similarity. They do not normally define the hypothesis space before the final identity decision.

Few-shot segmentation has progressed from support-conditioned dense prediction and prototype alignment to hypercorrelation, prior-guided enrichment, hierarchical correlation and optimal-transport matching[7–11]. Cross-domain variants further address appearance shifts through feature transformation, task-disruptive background suppression, iterative correspondence mining and representation recomposition[12–16]. These advances improve transferred appearance evidence, whereas Boundary-First changes the object representation on which that evidence acts.

Recent training-free systems sharpen semantic prompting in different ways. GF-SAM organises support–query correspondences as a graph, PR-MaGIC refines prompts through mask-decoder gradients, and INSID3 and FSSDINO exploit the semantic selectivity of DINOv3 features[17–20]. FSS-SAM3 instead places support and query images on a shared canvas and uses frozen SAM3 for visual-only or concept-prompted transfer[21]. These advances improve how an object is matched or prompted. UBET addresses an orthogonal question: which closed image regions should constitute admissible object hypotheses

before identity evidence chooses among them? Its contribution lies in the order and unit of inference, rather than in a new pretrained encoder or dataset-trained segmentation head.

The broader foundation-model literature has explored in-context segmentation, promptable recognition, open-world grounding and open-vocabulary dense prediction[22–28]. Self-supervised visual transformers provide the correspondence features used by many of these systems[4,29,30], while SAM2 and SAM3 extend promptable segmentation to videos and concepts[3,31]. Adaptation studies show that these foundation models can be repurposed for downstream segmentation[32–34]. UBET instead keeps the encoders frozen and tests whether closed-region construction can precede identity arbitration.

Classical contour grouping, region adjacency and superpixel methods take the complementary view that spatial coherence should be established explicitly. They reduce local ambiguity by organising pixels into regions, but generally do not address support-conditioned object identity. Modern few-shot methods, by contrast, are effective at transferring identity but commonly retain a semantic-to-mask formulation[35–38]. Entropy-regularised optimal transport provides a principled correspondence operator for structured matching[39]. Independent pixel decisions can also produce isolated foreground speckles far from the main object. Such disconnected predictions violate the spatial coherence expected of a single object, even when their local scores are high. This raises a distinct question: can few-shot segmentation construct a bounded object hypothesis before assigning its support identity?

Here we formulate this ordering as *Boundary-First segmentation*. Given a support image and mask, the query is first represented by a population of closed or groupable region hypotheses. Every hypothesis is evaluated through its interior, boundary and immediately exterior evidence, together with its enclosure structure. Support appearance may seed or rank atomic proposals, but closure is imposed before the final identity decision. The output is selected from bounded object hypotheses rather than assembled through unrestricted pixel assignment. This design is motivated by perceptual grouping and figure–ground organisation, which treat closure and enclosure as informative object-

level structure rather than as consequences of independent pixel labels.

We instantiate the principle in Unified Boundary Evidence Transport (UBET; Fig. 1). In the principal implementation, frozen SAM2 supplies class-agnostic masks and closed groups derived from prompts. Frozen DINOv2 transports identity evidence between support and query. DINOv2 maxima can seed prompted masks, but these masks are closed and grouped before final identity arbitration. Boundary evidence therefore defines competing object extents before semantic evidence selects or reconciles them. When a boundary-derived closed group and the strongest identity-scored region disagree, UBET applies a predeclared geometric conflict certificate rather than reverting to unrestricted pixel assignment. Because the boundary provider enters through a source-agnostic interface, the same candidate–evidence–selection order can also operate on visual-only SAM3 proposals, diffusion edges, depth edges, superpixels and conventional edge maps. The SAM3 instantiation changes only the proposal provider; DINOv2 evidence transport and the UBET certificate remain fixed.

Boundary-First denotes a family of segmentation systems rather than one fixed architecture. A boundary provider proposes and closes admissible object extents; an identity provider then ranks or reconciles those extents. Either provider can change while the inference order remains fixed. UBET is the strongest instantiation studied here, whereas the alternative-source variants are independent realisations of the same principle. Their purpose is not to claim that every edge detector is equivalent, but to test whether useful segmentation persists when the source of boundaries changes substantially.

We evaluate UBET with one-shot support on seven natural-image, video and medical benchmarks: FSS-1000, PASCAL-5^i, COCO-20^i, PerSeg, DAVIS, ISIC and Kvasir-SEG. Beyond aggregate mIoU, the experiments test the proposed mechanism. Independent boundary-source instantiations ask whether the principle persists across substantially different contour constructors. Matched UBET ablations remove boundary contrast or the conflict certificate while holding the candidate pool fixed. Frozen-backbone controls test whether the conclusion depends on one DINOv2 feature scale. Cross-dataset support transfer tests whether the same decision rule survives category and acquisition shifts with-

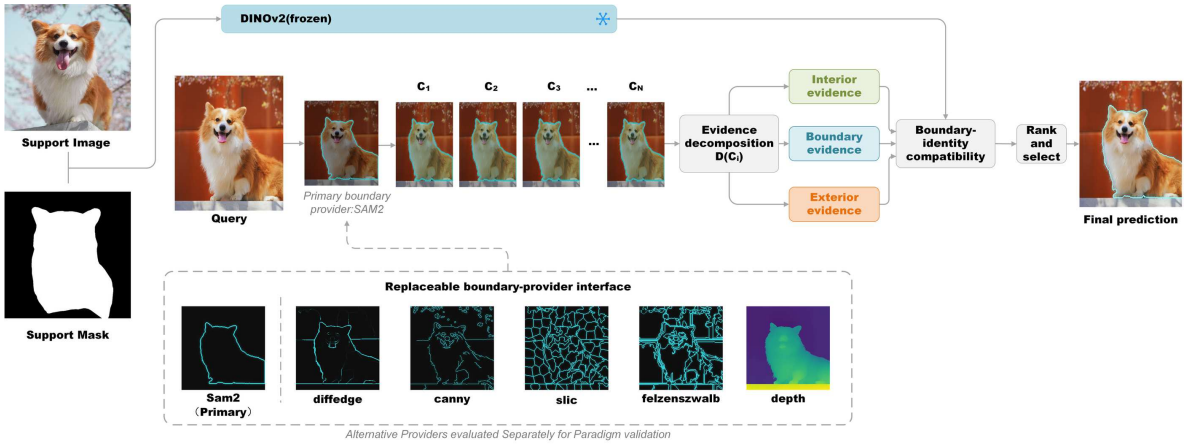


Figure 1: Unified Boundary Evidence Transport (UBET) implements Boundary-First one-shot segmentation as a continuous inference pipeline. An annotated support image defines object identity, while an unlabelled query is first converted into a population of closed region hypotheses. Each hypothesis is decomposed into interior, boundary and exterior evidence. Frozen DINOv2 features transport support identity to these bounded query regions, after which a compatibility certificate retains the closed group or reconciles it with the strongest identity hypothesis. The ranked decision yields a spatially coherent final mask. The lower interface shows that SAM2, diffusion and depth boundaries, superpixels and conventional edge maps can independently instantiate the same candidate–evidence–selection order.

out adaptation. Hypothesis-oracle and contour-quality analyses determine whether errors arise because a suitable region was never proposed or because an available region was not selected. An all-candidate ranking analysis then tests whether deployable evidence scores order post-hoc candidate quality within each query. Together, these experiments test three distinct claims. First, the ordering transfers across boundary sources and datasets. Second, its two evidence paths make separable contributions. Third, its remaining errors can be assigned to candidate availability or candidate selection.

2 Results

2.1 One-shot evaluation across category, object, video and medical domains

We evaluated UBET under one-shot protocols on seven benchmarks. PASCAL-5^i and COCO-20^i use the public four-fold, 1,000-episode-per-fold evaluation; FSS-1000 uses the complete 240-class, 2,400-image test set under the declared one-shot support protocol.

PerSeg, DAVIS, ISIC and Kvasir-SEG extend the evaluation to personalised objects, video, dermoscopy and endoscopy. The reported metric is foreground mean intersection-over-union (mIoU). SAM2 and DINOv2 remain frozen. Support-conditioned saliency is computed analytically from foreground and exterior-ring prototypes, without gradient optimisation or parameter fitting. Query annotations remain unavailable until post-hoc evaluation. Headline scores and all mIoU ablations follow each benchmark's evaluation resolution. The boundary F-score is a separate diagnostic computed on a common 256-by-256 working grid. Mechanistic ablations and diagnostics use the SAM2 implementation unless a different boundary provider is stated explicitly.

UBET (SAM2) obtains 73.34% mIoU on PASCAL-5^i under the matched four-fold protocol (Table 1). Its COCO-20^i score is 59.34%. The controlled UBET (SAM3) instantiation reaches 83.19% and 57.70%, respectively. UBET (SAM3) has the highest observed score on every PASCAL-5^i fold, whereas UBET (SAM2) has the highest score on three of four COCO-20^i folds. The SAM2 COCO-20^i four-fold mean is 0.70 mIoU points above GF-SAM (59.34 versus 58.64). Across the broader evaluation, the principal SAM2 configuration reaches 87.24% on FSS-1000, 94.68% on PerSeg and 80.83% on DAVIS. It obtains 65.52% on ISIC and 70.14% on Kvasir-SEG (Table 2).

Matched qualitative episodes complement the aggregate comparison across all seven domains (Fig. 2). Every method receives the same one-shot support–query pair within each row. Across the displayed natural, video and medical images, UBET preserves finer contours, suppresses isolated foreground speckles and reduces leakage into neighbouring structures. Its predictions consequently form more coherent closed regions, including under occlusion and weak local contrast.

Performance varied by provider and domain. UBET (SAM2) had the highest observed FSS-1000 and DAVIS scores, FSS-SAM3 had the highest PerSeg score, and UBET (SAM3) had the highest medical scores. On ISIC and Kvasir-SEG, UBET (SAM3) exceeded direct FSS-SAM3 by 5.9 and 14.0 mIoU points, respectively.

Table 1: Matched one-shot four-fold evaluation on PASCAL-5^i and COCO-20^i. Values are foreground mIoU percentages over 1,000 episodes per fold under identical fold definitions and one annotated visual support example. Mean ± s.d. summarises the four fold-specific class-aggregated scores. FSS-SAM3 uses released visual-only inference. Bold and underlined entries denote the best and second-best values, respectively.

a PASCAL-5^i

Method	Fold 0	Fold 1	Fold 2	Fold 3	Mean ± s.d.
PerSAM[5]	50.18	53.35	45.27	39.98	47.20 ± 5.85
PerSAM-F[5]	46.20	51.03	46.69	38.31	45.56 ± 5.30
FSSDINO[20]	51.46	62.56	58.82	51.46	56.08 ± 5.54
PR-MaGIC[18]	66.65	67.36	63.17	65.54	65.68 ± 1.83
INSID3[19]	63.77	73.14	67.96	68.42	68.32 ± 3.83
Matcher[6]	67.66	70.29	66.05	67.69	67.92 ± 1.75
GF-SAM[17]	70.61	75.78	68.32	72.40	71.78 ± 3.15
ABCDFSS[40]	35.17	43.83	40.97	36.51	39.12 ± 4.00
FSS-SAM3[21]	<u>76.00</u>	<u>81.14</u>	<u>82.27</u>	<u>84.71</u>	<u>81.03 ± 3.67</u>
UBET (SAM2)	74.69	77.02	68.58	73.08	73.34 ± 3.56
UBET (SAM3)	**81.15**	**82.93**	**83.08**	**85.61**	**83.19 ± 1.84**

b COCO-20^i

Method	Fold 0	Fold 1	Fold 2	Fold 3	Mean ± s.d.
PerSAM[5]	22.62	23.74	21.40	20.79	22.14 ± 1.31
PerSAM-F[5]	21.85	23.59	23.13	22.34	22.73 ± 0.78
FSSDINO[20]	38.72	41.33	40.12	42.44	40.65 ± 1.60
PR-MaGIC[18]	50.46	53.65	56.32	53.13	53.39 ± 2.40
INSID3[19]	52.34	55.41	54.27	53.82	53.96 ± 1.27
Matcher[6]	51.48	52.76	51.22	52.72	52.05 ± 0.81
GF-SAM[17]	56.56	**60.96**	<u>59.33</u>	<u>57.71</u>	<u>58.64 ± 1.92</u>
ABCDFSS[40]	22.02	25.40	24.24	24.35	24.00 ± 1.42
FSS-SAM3[21]	56.65	54.26	53.80	55.32	55.01 ± 1.27
UBET (SAM2)	<u>57.64</u>	<u>60.13</u>	**60.11**	**59.46**	**59.34 ± 1.17**
UBET (SAM3)	**58.54**	58.29	58.85	55.12	57.70 ± 1.74

2.2 Boundary-First arbitration transfers across datasets

We next changed the dataset from which the one-shot support was drawn while leaving the query benchmark and inference rule unchanged. COCO-to-PASCAL and PASCAL-to-COCO evaluate category transfer over shared semantic classes; Kvasir-SEG-to-ETIS evaluates transfer between independent endoscopic collections. Every method receives the same matched one-shot information, and none is adapted to the target queries.

The principal UBET (SAM2) configuration obtains 51.55% mIoU with COCO support and PASCAL queries. The reverse direction yields 45.40%, and Kvasir-SEG support

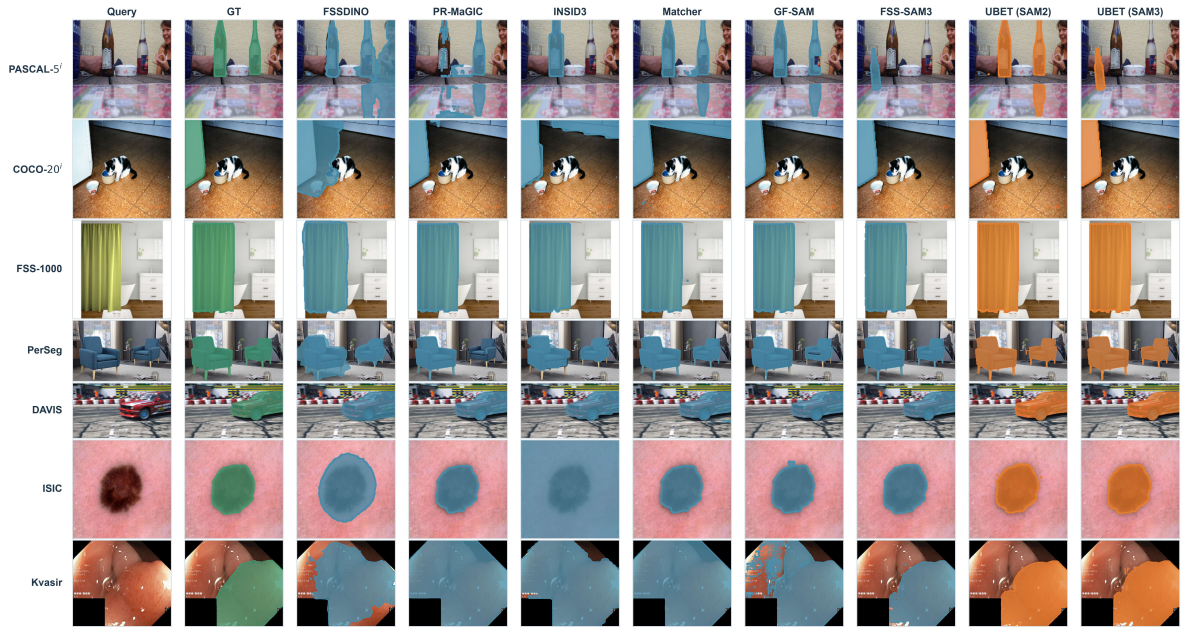


Figure 2: Image-grounded comparison across seven one-shot segmentation benchmarks. The query and ground truth are followed by released-code predictions from FSSDINO, PR-MaGIC, INSID3, Matcher, GF-SAM and visual-only FSS-SAM3, followed by the SAM2 and SAM3 UBET instantiations. Every method receives the same support–query pair within a row. Coloured foreground overlays and contours reveal incomplete extent, leakage, isolated predictions and boundary closure.

Table 2: One-shot segmentation across natural-image, personalised-object, video and medical benchmarks. Values are foreground mIoU percentages on the complete query partitions under the declared protocols. Comparator provenance and adapter status are documented in Methods. Bold and underlined entries denote the best and second-best values, respectively.

Method	FSS-1000	PerSeg	DAVIS	ISIC	Kvasir-SEG
PerSAM[5]	72.3	83.3	61.7	18.9	17.4
PerSAM-F[5]	75.1	91.6	65.7	43.7	36.2
Matcher[6]	85.1	86.9	72.6	50.7	36.0
GF-SAM[17]	<u>85.5</u>	87.2	70.6	44.2	60.4
INSID3[19]	83.1	90.1	73.5	52.2	43.9
FSSDINO[20]	75.2	82.1	60.2	48.8	41.4
PR-MaGIC[18]	84.7	94.1	<u>79.6</u>	54.0	56.8
ABCDFSS[40]	69.2	82.8	56.4	44.2	55.1
FSS-SAM3[21]	62.2	**95.1**	71.6	<u>66.8</u>	65.7
UBET (SAM2)	**87.2**	<u>94.7</u>	**80.8**	65.5	<u>70.1</u>
UBET (SAM3)	71.3	93.6	73.3	**72.7**	**79.7**

with ETIS queries yields 67.40% (Table 3). Replacing the proposal provider with SAM3 changes these scores to 42.63%, 46.71% and 73.61%, respectively. The released visual-only

FSS-SAM3 comparator reaches 35.77%, 30.97% and 73.80%. UBET (SAM2) is strongest in the first direction, whereas UBET (SAM3) leads the reverse direction. FSS-SAM3 and UBET (SAM3) are nearly tied for endoscopic transfer. This provider-dependent ranking confirms that cross-dataset transfer remains difficult, although both UBET instantiations improve on direct FSS-SAM3 in the two natural-image directions. These experiments test transfer of each complete inference rule; they do not use target-query annotations or target-dataset optimisation.

Table 3: Cross-dataset one-shot segmentation without target-dataset adaptation. Natural-image transfer uses 100 target queries per shared-category direction; Kvasir-SEG-to-ETIS uses all 182 ETIS queries. Values are foreground mIoU percentages under matched visual-support protocols. Bold and underlined entries denote the best and second-best values, respectively.

Method	COCO → PASCAL	PASCAL → COCO	Kvasir → ETIS
PerSAM[5]	32.00	30.22	15.35
PerSAM-F[5]	35.88	32.76	34.87
FSSDINO[20]	41.49	33.80	30.12
PR-MaGIC[18]	35.69	36.48	29.61
INSID3[19]	46.70	39.30	48.60
Matcher[6]	39.91	42.04	34.30
GF-SAM[17]	<u>50.75</u>	43.91	50.28
ABCDFSS[40]	32.13	26.81	27.81
FSS-SAM3[21]	35.77	30.97	**73.80**
UBET (SAM2)	**51.55**	<u>45.40</u>	67.40
UBET (SAM3)	42.63	**46.71**	<u>73.61</u>

Table 4: Independent Boundary-First implementations using alternative boundary sources. FSS-1000 uses the complete 240-class, 2,400-image test set, PerSeg uses 176 queries and DAVIS uses a common 300-frame, 30-sequence protocol. Values are foreground mIoU percentages; the best and second-best values in each column are shown in bold and underlined, respectively.

Boundary source	FSS-1000	PerSeg	DAVIS
Diffusion edge + depth edge	**87.33**	**91.22**	77.46
Diffusion edge	83.35	<u>91.19</u>	**79.37**
Depth edge	79.01	88.57	<u>77.70</u>
SLIC superpixels	85.61	91.07	74.50
Canny edges	85.12	87.80	73.03
Felzenszwalb regions	<u>86.04</u>	90.97	73.92

2.3 Boundary-First inference transfers across boundary sources

The principal UBET implementation uses SAM2 to populate the query candidate set, but the proposed inference order is not tied to this provider. We therefore built six independent Boundary-First systems that construct closed hypotheses from alternative boundary sources and then apply the same sequence of contour closure, region verification and support-conditioned arbitration. Diffusion edges, depth edges and their combination provide learned alternatives; the depth maps come from the frozen Depth Anything V2 Small checkpoint. SLIC, Canny and Felzenszwalb provide conventional superpixel, edge and graph-region partitions[36,41–44]. These are separate Boundary-First implementations rather than degraded variants of the principal UBET model.

The independent alternative-source systems yield consistent segmentation on FSS-1000, PerSeg and DAVIS (Table 4). Combined diffusion and depth evidence reaches 87.33, 91.22 and 77.46% mIoU, respectively, whereas SLIC reaches 85.61, 91.07 and 74.50%. Diffusion-only and depth-only variants also remain competitive, although their relative ranking changes by domain. Felzenszwalb regions provide a further conventional control, reaching 86.04, 90.97 and 73.92% mIoU on FSS-1000, PerSeg and DAVIS, respectively. Canny reaches 85.12, 87.80 and 73.03% on the same three domains. These results do not imply that all boundary sources are equally informative. Instead, they show that the Boundary-First decision rule remains effective when candidate contours are obtained in substantially different ways. Provider-specific construction details are given in Supplementary Methods 1.

Figure 3 links these aggregate results to the construction of independent Boundary-First systems. In an FSS-1000 episode distinct from the main framework example, fused diffusion–depth evidence, DiffusionEdge, depth boundaries, SLIC, Canny and Felzenszwalb produce visibly different partitions. The shared Boundary-First interface nevertheless converts each source into a coherent prediction. The resulting masks preserve closed object extent despite marked differences in the underlying partitions, consistent with the quantitative results in Table 4.

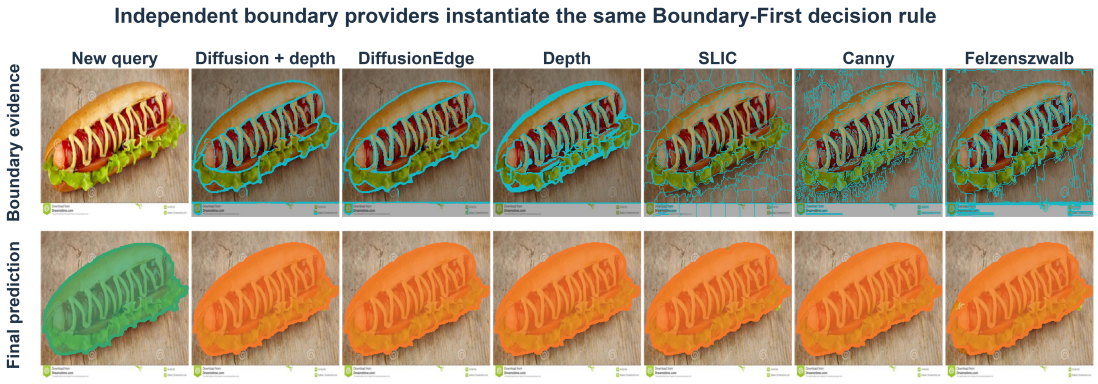


Figure 3: Image-grounded transfer across independent boundary providers. The upper row shows the query and the evidence supplied by fused diffusion–depth cues, DiffusionEdge, depth boundaries, SLIC, Canny and Felzenszwalb. The lower row shows retrospective ground truth and the corresponding final predictions. Each provider independently constructs its candidate regions, while all six systems share the same closed-region verification and identity-arbitration order. Distinct partitions converge on coherent object extent under the shared decision rule.

2.4 Ablation experiments isolate conflict repair and identity evidence

The locked UBET decision begins with the support-conditioned closed-group hypothesis. It invokes an identity-defined intersection only when this group and the strongest identity-scored region have IoU below 0.25. We recomputed every decision control from identical frozen candidate evidence, so no ablation regenerates proposals or changes support evidence. Relative to retaining the boundary group alone, the full certificate improves PerSeg, DAVIS, ISIC and Kvasir-SEG by 0.97, 1.56, 3.81 and 7.02 mIoU points, respectively (Table 5a). Removing boundary contrast causes larger losses of 2.65, 3.42, 2.23 and 5.47 points on the same datasets. These population-level gains support the mechanism formalised in Proposition 3: conflict-conditioned reconciliation is beneficial on average, although its sufficient inequality need not hold for every query.

We then reversed the decision order while retaining the same frozen evidence. The Boundary-First order outperforms the identity-first order by 2.60 points on PerSeg, 3.97 on DAVIS, 3.20 on ISIC, 4.50 on Kvasir-SEG and 1.60 on FSS-1000 (Table 5b). The sign is consistent across object, video, medical and category-level domains. Together with the component controls, this result supports the ordering itself rather than only the presence

of an additional boundary score. Because closure and identity pruning generally do not commute (Proposition 2), this matched reversal directly tests the proposed decision order while holding the evidence sources fixed.

Table 5: Mechanism ablation and failure decomposition. **a**, Conflict-certificate ablation on identical candidate pools. **b**, Decision-order reversal on frozen evidence. **c**, Candidate-ceiling and geometric diagnostics; Oracle is the best post-hoc IoU in the retained hypothesis family, Gap is Oracle minus Selected, and the BF score uses a two-pixel tolerance. Values are complete-partition percentages; query annotations are accessed only after prediction.

a Boundary and conflict-certificate components

Dataset	n	No contrast	Boundary only	Full UBET	Gain vs boundary
PerSeg	176	92.03	93.71	**94.68**	0.97
DAVIS	1,969	77.41	79.27	**80.83**	1.56
ISIC	650	63.29	61.71	**65.52**	3.81
Kvasir-SEG	200	64.67	63.12	**70.14**	7.02

b Decision-order intervention

Dataset	n	Boundary-First	Identity-first	Difference
PerSeg	176	**94.70**	92.10	+2.60
DAVIS	1,969	**80.80**	76.83	+3.97
ISIC	650	**65.50**	62.30	+3.20
Kvasir-SEG	200	**70.10**	65.60	+4.50
FSS-1000	2,400	**87.20**	85.60	+1.60

c Candidate ceiling and geometric diagnostics

Dataset	n	Selected	Oracle	Gap	BF score	Mean candidates
PerSeg	176	94.68	95.34	0.66	86.97	29.9
DAVIS	1,969	80.83	86.83	6.00	81.01	37.9
ISIC	650	65.52	77.59	12.07	37.77	24.4
Kvasir-SEG	200	70.14	87.13	16.99	45.54	28.3

Additional complete-path controls and frozen DINOv2 feature-scale comparisons are reported in Supplementary Tables 2 and 3; sensitivity to the shared conflict threshold is reported in Supplementary Table 4. One decision path, backbone configuration and threshold are used across all datasets.

The image-grounded controls in Fig. 4 clarify what changes between fixed component paths. On DAVIS and Kvasir-SEG, semantic-only selection and unconditional intersection retain only a discriminative object fragment, whereas the closed boundary hypothesis

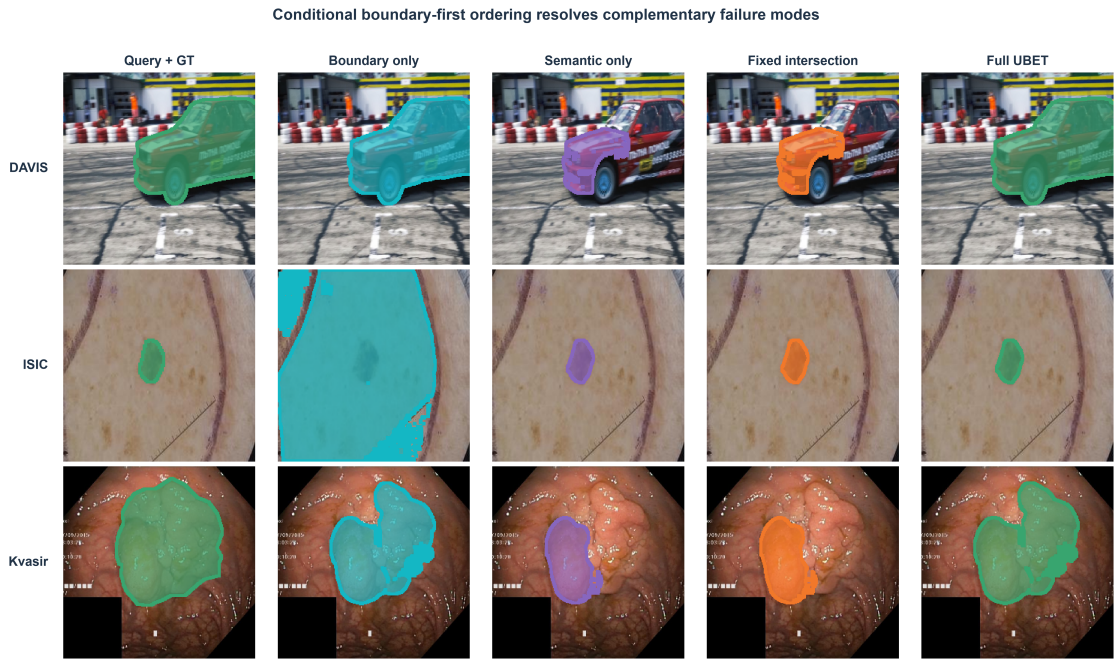


Figure 4: Complementary failure modes of fixed decision paths. Three real episodes compare boundary-only, semantic-only, unconditional-intersection and full conditional inference using identical frozen evidence. Boundary evidence preserves object extent in the DAVIS and Kvasir-SEG examples, whereas identity evidence repairs an overextended boundary hypothesis in ISIC. Full UBET selects the appropriate operation from evidence compatibility, preserving closed extent when the streams agree and suppressing unsupported regions when they diverge.

preserves the complete extent. The ISIC episode exposes the complementary case: a weak boundary proposal expands over most of the image, while identity evidence and conflict repair isolate the lesion. UBET therefore conditions the use of intersection on measured incompatibility rather than treating either evidence source as universally reliable. These images illustrate the measured mechanisms; complete-partition effects are reported in Table 5.

We next tested whether the deployable evidence score provides a valid within-query ranking of candidate quality. The audit retained every closed candidate from all 2,400 FSS-1000 queries in the complete 240-class test set (74,061 candidates) and compared its fused evidence score with post-hoc candidate IoU. Because raw scores are query dependent, all rank associations were calculated within query. This directly tests the query-local ordinal premise of the region verifier. The pooled within-query Spearman rank correlation coefficient was 0.876 (query-cluster bootstrap 95% confidence interval, 0.871–0.881;

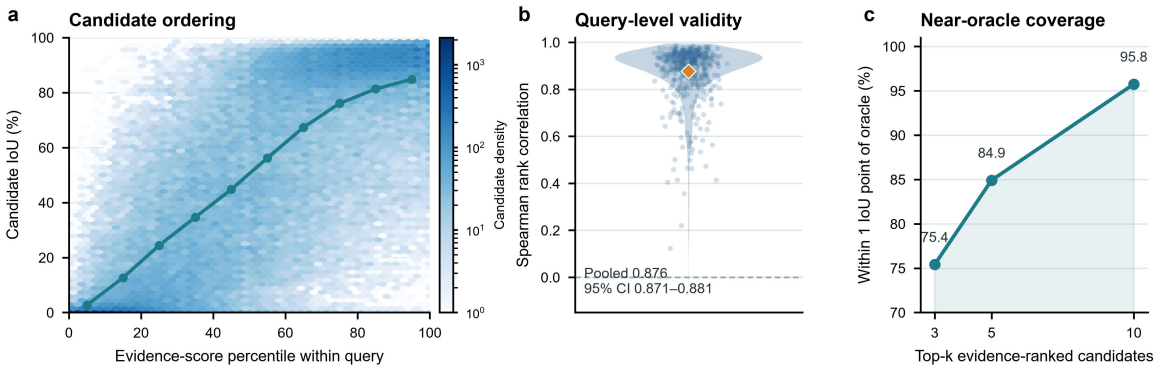


Figure 5: Query-local evidence-score validity across all retained candidates. **a**, Density of candidate IoU across within-query evidence-score percentiles for 74,061 candidates from all 2,400 FSS-1000 queries; the line and shading show query-aggregated percentile-bin means and query-cluster bootstrap 95% confidence intervals. **b**, Distribution of query-level Spearman rank correlation coefficients; the orange diamond and vertical interval mark the pooled within-query coefficient and its query-cluster bootstrap 95% confidence interval. **c**, Near-oracle coverage among the top three, five and ten evidence-ranked hypotheses. IoU is used only for this post-hoc audit and is unavailable to the selector.

Fig. 5), and 99.875% of query-level coefficients were positive. The best candidate among the top three evidence-ranked hypotheses was within one IoU point of the full candidate oracle for 75.42% of queries. Coverage increased to 84.92% for the top five and 95.75% for the top ten. These results establish a strong query-local ordering and show that a compact evidence-ranked shortlist preserves near-oracle candidates. Raw scores are not interpreted as globally calibrated probabilities.

2.5 Candidate availability and selection limits

The candidate oracle measures the best mask already present in the retained hypothesis family. Oracle minus Selected measures selection regret within that family. This gap diagnoses ranking, whereas the oracle ceiling reflects candidate availability. Across the diagnostic partitions, mean selection regret is 0.66 points on PerSeg, 6.00 on DAVIS, 12.07 on ISIC and 16.99 on Kvasir-SEG (Table 5). PerSeg is close to its attainable ceiling. On ISIC and Kvasir-SEG, substantially better retained hypotheses are often available than the region selected by the deployed identity arbitration. All oracle quantities are computed from ground truth only after inference and are unavailable to UBET.

The boundary F-score provides a complementary geometric view without repeating

several closely related surface metrics in the main text. PerSeg reaches a BF score of 86.97%, DAVIS reaches 81.01%, and ISIC and Kvasir-SEG have substantially weaker contour scores. Larger oracle gaps co-occur with weaker BF scores, while the retained high-IoU candidates show that arbitration remains a separate source of error.

2.6 Accuracy, footprint and inference cost

We profiled the released-code comparators under serial execution after one-time model initialisation. Total parameters count all instantiated pretrained weights because reporting only trainable parameters would hide the dominant foundation-model footprint. The accuracy column in Supplementary Table 7 is the descriptive arithmetic mean of the seven dataset-specific headline scores. It summarises breadth but does not replace the dataset-level comparisons in Tables 1 and 2. PR-MaGIC uses its released SAM ViT-H and DINOv2 ViT-L/14 stack. INSID3 uses the repository-default frozen DINOv3-L/16 backbone[30]; we count 303.15 million parameters by instantiating the official architecture and record an end-to-end inference time of 0.7331 seconds per query. FSS-SAM3 obtains a seven-dataset mean of 71.06% with 840.51 million parameters and a measured latency of 0.698 seconds per query.

The SAM3 and SAM2 UBET instantiations have the highest seven-dataset means, 75.95% and 75.87%, respectively. The SAM2 model exceeds PR-MaGIC by 6.12 points, and its 615.40 million parameters are fewer than PerSAM's 641.09 million and the 945.46 million parameter SAM–DINO stacks used by Matcher, GF-SAM and PR-MaGIC. The footprint remains larger than those of the DINO-only and ResNet-50 comparators (Supplementary Table 7). UBET requires 2.481 seconds per query in the common-hardware profile. It is slower than FSSDINO, FSS-SAM3, GF-SAM, INSID3 and PerSAM, but faster than ABCDFSS, Matcher, PerSAM-F and PR-MaGIC.

Stage profiling locates 40.79% of UBET time in SAM2 proposal generation and 37.17% in identity transport (Supplementary Table 8). Together, these stages account for 77.96% of latency. The final certificate contributes approximately 0.01%. The locked configuration samples four automatic SAM2 points per side and retains 39.47 unique hypotheses

per query on average in this profile. The SAM3 implementation contains 840.51 million proposal-model parameters and 1,231.46 million parameters after including the frozen DINOv2 evidence encoders. Its measured FSS-1000 latency is 1.8146 seconds per query.

3 Discussion

The central result is that closed regions can serve as the decision units of one-shot segmentation before identity arbitration. This ordering persists in independent implementations based on diffusion, depth, superpixels, Canny and Felzenszwalb regions. Boundary-source transfer therefore tests the organisation of inference independently of any one proposal foundation model. The seven-benchmark evaluation tests the locked UBET implementation across category, object, video and medical domains, while cross-dataset support adds a stricter distribution shift. Together, the principal model and six alternative-source systems provide a family-level test of Boundary-First segmentation rather than an ablation of a single proposal backbone.

The SAM3 experiment provides a complementary foundation-model control. Visual-only FSS-SAM3 is already a strong direct segmenter, yet applying the same SAM3 proposal interface inside UBET raises the PASCAL-5^i mean from 81.03% to 83.19%. The corresponding COCO-20^i result, 57.70%, remains below the 59.34% SAM2 UBET result. Thus, SAM3 is not presented as a universal replacement for SAM2. Instead, the two rows show that the closed-region evidence and arbitration rule can transfer to a newer proposal foundation while preserving a visual-only, training-free formulation.

The UBET controls locate the contribution more precisely. The conflict certificate reconciles a boundary-defined group with an independently identity-scored region. Relative to the uncorrected boundary group, the full decision improves all four tested domains, with the largest gains on ISIC and Kvasir-SEG. The result supports a bounded mechanism claim: boundary hypotheses define the initial object extent, while semantic evidence rejects unsupported parts when the two representations conflict. Proposition 3 makes this claim precise: intersection improves overlap only when the false-positive area removed by

identity evidence compensates for any target area that is discarded. Conditional repair is therefore structurally preferable to unconditional intersection.

The decision-order intervention tests more than the presence of two useful signals. As formalised by Proposition 2, closing a contour and pruning its atoms by identity evidence are generally non-commutative operations. The consistent advantage of Boundary-First processing across five domains shows that retaining the complete geometric hypothesis before identity arbitration is consequential in practice.

Hypothesis-oracle analysis further separates errors in contour availability from errors in selection. Proposition 1 makes this distinction exact: the oracle ceiling measures what the retained hypothesis family can represent, whereas the oracle gap measures selection regret within that family. PerSeg is close to its ceiling, while the larger medical-domain gaps identify selection headroom for morphologically variable lesions. Better boundary providers would raise a low ceiling when no suitable contour exists; better correspondence or additional semantic evidence would reduce selection regret when it does. The two stages can therefore be improved and evaluated independently without changing the Boundary-First order.

The candidate-level ranking audit complements this oracle decomposition. The fused evidence score is strongly monotonic with candidate IoU within query, and a short evidence-ranked list recovers a near-oracle hypothesis for most queries. The positive association in 99.875% of queries and 95.75% top-ten coverage support the proposed ordering of closed-region hypotheses before final identity selection. This conclusion is limited to the FSS-1000 audit and does not imply probability calibration across queries or datasets. The ordinal claim is more directly testable than treating evidence scores as absolute IoU estimates.

Appearance and boundary evidence remain complementary. Single-backbone and alternate-path controls vary in their relative ranking across domains, so we retain one predeclared B+L visual configuration for every dataset rather than selecting a backbone or decision path per benchmark. Region-specific perturbations further show that contour and interior evidence provide distinct information (Supplementary Tables 5 and 6).

UBET also inherits errors from both frozen evidence providers. Sparse SAM2 sampling can omit thin or weakly bounded targets, while pretrained DINOv2 correspondences can favour a visually related distractor or only the most diagnostic object part. The fixed conflict threshold is deliberately shared across datasets. Although the full route is strongest among the evaluated controls, this does not establish universal optimality over every possible selector. The method is training-free: it performs no dataset-level task-specific training, episodic gradient optimisation or checkpoint adaptation. All support-conditioned evidence is computed analytically, and no parameters are retained between episodes. UBET nevertheless relies on large pretrained models and is not a lightweight segmenter.

Resource profiling refines this limitation. The full UBET footprint is smaller than the SAM–DINO stacks used by several comparators. Proposal generation and identity transport nevertheless account for approximately 78% of its latency. SAM2-Tiny plus DINOv2-L reduces this cost, but the complete COCO fold records a 2.34-point loss despite a favourable first-100 diagnostic. Checkpoint scaling therefore defines a practical accuracy–resource trade-off rather than a lossless replacement. The full configuration remains the basis of every headline comparison.

The boundary-source variants support ordering portability but do not all share the same computational cost or candidate distribution. Their role is to test whether the Boundary-First order survives source replacement, not to claim that every boundary detector is interchangeable. Hypothesis-oracle scores are post-hoc diagnostic ceilings, not deployable predictions.

Boundary-First is a reusable segmentation paradigm rather than a fixed collection of frozen components. Its defining commitment is procedural: construct a bounded object hypothesis first, then determine whether the enclosed region matches the support identity. The error decomposition suggests two direct extensions. A task-trained or self-learning boundary provider could raise the candidate ceiling by recovering support-relevant closed extents. A block- or region-matching model could independently reduce selection regret among plausible hypotheses. Stronger region correspondence, including

carefully controlled language evidence, offers a further route, although no reported result uses multimodal input. The present study does not estimate the gains of these future modules.

In summary, UBET operationalises Boundary-First segmentation as closed-region hypothesis construction followed by support-conditioned identity arbitration. Its performance across seven benchmarks, persistence under six alternative boundary providers and targeted decision-order controls show that this ordering is both effective and portable. The resulting stage-wise formulation provides a measurable foundation for future trainable or self-learning systems that improve contour availability and identity selection independently.

4 Methods

4.1 Datasets and evaluation protocol

All experiments use one annotated support image and a disjoint query image. FSS-1000, PASCAL-5^i and COCO-20^i evaluate category-level transfer[1,7,45]. PerSeg evaluates personalised objects[5], and DAVIS evaluates annotated video objects[46]. ISIC evaluates skin lesions[47]; Kvasir-SEG and ETIS evaluate polyps[48,49]. PASCAL-5^i and COCO-20^i use the public seed-zero four-fold manifests with 1,000 episodes per fold. FSS-1000 uses the complete 240-class, 2,400-image test set and a fixed one-shot support–query manifest shared by every evaluated method. PerSeg and DAVIS use all held-out object groups and validation sequences, respectively. ISIC and Kvasir-SEG use 650 and 200 held-out queries. Empty or failed predictions are retained as zero-IoU cases. Support selection is fixed before evaluation. All methods share the resulting episode manifests, and no support image is selected according to query performance. Query masks remain unavailable until prediction, except in explicitly labelled post-hoc diagnostics.

Cross-dataset transfer uses shared natural-image categories for COCO-to-PASCAL and PASCAL-to-COCO, with 100 target queries in each direction. Kvasir-SEG-to-ETIS uses all 182 ETIS target queries. The one-shot support protocol is applied identically

to every method, and no target-query annotation, target-dataset training or persistent checkpoint update is permitted. Comparator methods may retain their declared support-only episodic operations, but no state produced in one episode is carried into another. UBET performs no episodic fitting or gradient optimisation.

The metric is foreground mean intersection-over-union (mIoU), also commonly denoted J. FSS-1000, PASCAL-5^i and COCO-20^i use their class-aggregated protocols; PerSeg and DAVIS aggregate over object groups or sequences; the two medical datasets use query means. No Dice score is used. Region construction and evidence scoring operate on a 256-by-256 working grid. DINOv2 identity inputs use 518-by-518 pixels, and SAM2 retains its native predictor preprocessing. For PASCAL-5^i and COCO-20^i, predictions are mapped through the stored query transform before metric accumulation.

4.2 Frozen visual evidence providers

The locked UBET implementation combines SAM2 Hiera-L with frozen DINOv2 ViT-B/14 and ViT-L/14 encoders. SAM2 experiments use the SAM 2.1 Hiera-L checkpoint `sam2.1_hiera_large`, referred to as SAM2 throughout. Automatic proposal generation uses four points per side, 64 points per batch and no crop layer. Predicted-IoU and stability thresholds are 0.75 and 0.85, box non-maximum suppression is 0.7, and the minimum mask area is 100 working-grid pixels. DINOv2 patch and class tokens are normalised within each encoder; B/L evidence is combined in the reciprocal region-matching and dense-localisation routes. All networks remain in evaluation mode and inference uses half-precision SAM2 execution. The SAM2 and DINOv2 parameters receive no gradient update, dataset-level optimisation or checkpoint adaptation. Query masks are never used before evaluation.

For the controlled SAM3 instantiation, the SAM2 proposal generator is disabled and replaced by the released visual-only FSS-SAM3 pairwise canvas interface with the official `sam3.pt` checkpoint. Text prompts and class names are disabled. A confidence threshold of 0.30 is fixed globally for UBET (SAM3); the DINOv2 B+L identity representation, candidate evidence, conflict threshold and final selection rule are otherwise unchanged.

The FSS-SAM3 comparator uses the released visual-only inference and parameters without UBET evidence transport. Neither SAM3 route updates a model weight or retains episode-specific state between queries.

4.3 Problem formulation

We consider one-shot segmentation with a support pair $\mathcal{S} = \{(x_s, m_s)\}$, where x_s is the support image and m_s its binary foreground mask. Given an unlabelled query image q, the goal is to predict a binary mask $\hat{m}_q$ for the object represented by the support mask. Let $\phi(\cdot)$ denote the frozen DINOv2 feature extractor and let $\mathcal{B}(q; \mathcal{S})$ denote a boundary-based proposal operator that may use support-conditioned seeds. Unlike pixel-wise decoding, Boundary-First inference first restricts the output space to a finite family of closed hypotheses, $\mathcal{H}_{\mathcal{B}}(q)$. Its decision architecture can be written as the chain factorisation

$$p(h, z = 1 \mid q, \mathcal{S}) = p_{\mathcal{B}}(h \mid q, \mathcal{S}) p_{\mathcal{I}}(z = 1 \mid h, q, \mathcal{S}), \qquad h \in \mathcal{H}_{\mathcal{B}}(q),$$

where the first factor constructs geometrically admissible regions and the second evaluates whether a region carries the support identity. This factorisation is not an independence assumption; it specifies that identity is decided conditionally on an already bounded region. UBET instantiates the factorisation through candidate construction, structural closure, identity arbitration and conflict resolution:

$$\mathcal{R}_q = \mathcal{B}(q; \mathcal{S}), \qquad g_q = \Gamma(\mathcal{R}_q; \mathcal{S}, q),$$

$$r_s = \arg\max_{r \in \mathcal{R}_q} \Psi(r; \mathcal{S}, q), \qquad \hat{m}_q = \mathcal{C}(g_q, r_s).$$

Here $\mathcal{R}_q \subseteq \mathcal{H}_{\mathcal{B}}(q)$ contains closed or groupable hypotheses, Γ selects the structurally preferred group and Ψ is a support-conditioned region verifier. The certificate $\mathcal{C}$ either retains g_q or reconciles it with r_s. Support evidence may seed proposal construction, but semantic evidence only ranks or reconciles bounded regions; it does not independently label query pixels.

4.4 Theoretical rationale and error decomposition

The formulation yields four testable consequences.

Proposition 1 (stage-wise error decomposition). Let m_q be the unknown query target and let $\mathcal{R}_q(\mathcal{B})$ be the closed candidate set induced by boundary operator $\mathcal{B}$. The certificate-augmented hypothesis family is

$$\mathcal{H}_q = \mathcal{R}_q \cup \{\mathcal{C}(g_q, r_s)\}.$$

Define the attainable ceiling and selection regret as

$$O_q = \max_{h \in \mathcal{H}_q} J(h, m_q), \qquad \Delta_q = O_q - J(\hat{m}_q, m_q).$$

Because $\hat{m}_q \in \mathcal{H}_q$, the total overlap error decomposes exactly as

$$1 - J(\hat{m}_q, m_q) = \underbrace{1 - O_q}_{\text{candidate-generation error}} + \underbrace{\Delta_q}_{\text{selection regret}}.$$

This identity separates limitations of the boundary provider from those of the verifier. Defining $\epsilon_{\mathcal{B}} = \mathbb{E}_q[1 - O_q]$ and $\epsilon_{\mathcal{I}} = \mathbb{E}_q[\Delta_q]$ gives the population relation

$$1 - \mathbb{E}_q[J(\hat{m}_q, m_q)] = \epsilon_{\mathcal{B}} + \epsilon_{\mathcal{I}}.$$

The first term measures representational error in the retained closed-region family; the second measures arbitration error within that family. A high ceiling with large Δ_q calls for better identity arbitration, whereas a low ceiling cannot be repaired by reranking the same pool.

Proposition 2 (order is generally non-commutative). Let G close and group boundary atoms and let P_Ψ prune atoms using identity evidence. In general,

$$P_\Psi(G(\mathcal{R}_q)) \neq G(P_\Psi(\mathcal{R}_q)).$$

If a weakly matched atom is required to complete a true contour, pruning before grouping removes information that closure cannot subsequently recover. The Boundary-First order preserves the complete geometric hypothesis until its identity is evaluated; the matched order-reversal experiment tests this claim with the same evidence sources.

Proposition 3 (when conflict intersection improves overlap). Let g be the closed structural hypothesis, r the identity region and m the target. Write $A = |m|$, $a = |g \cap m|$, $b = |g \setminus m|$, $x = |(g \setminus m) \setminus r|$ and $y = |(g \cap m) \setminus r|$. Direct expansion of intersection-over-union gives

$$J(g \cap r, m) \geq J(g, m) \quad \Longleftrightarrow \quad ax \geq y(A + b).$$

Thus intersection is beneficial when the identity region removes enough unsupported area (x) to compensate for discarded target area (y). This is why UBET invokes intersection only after a conflict certificate rather than as an unconditional refinement; the condition is sufficient for an individual query, whereas the ablation tests its aggregate effect.

Proposition 4 (geometric sensitivity of region overlap). Let an ideal planar target m have area A and perimeter P, and suppose a predicted contour lies within a width-d tubular neighbourhood of the true contour. Under the usual regularity condition for such a neighbourhood, the symmetric-difference area obeys the conservative bound $\delta_d \leq 2dP + \pi d^2$. Whenever $\delta_d < A$,

$$J(r, m) \geq \frac{A - \delta_d}{A + \delta_d}.$$

The bound exposes the perimeter-to-area ratio P/A as the relevant geometric condition: the same contour displacement causes a larger overlap penalty for small, thin or highly irregular targets. It complements Proposition 1 by showing how boundary quality controls the attainable candidate ceiling.

Finally, the deployable evidence score is treated as a query-local ordinal surrogate, not as a calibrated estimate of IoU. Its required property is that better candidates tend to receive better ranks within the same query. We test this premise directly through

candidate-level rank correlation, top-k oracle coverage and selection regret; no ground-truth utility is used at inference.

4.5 Overview of Boundary-First inference

UBET follows the same order in every episode. First, SAM2 produces an automatic population of closed masks, and support-conditioned prompts produce additional atomic masks and connected closed-group hypotheses. Second, two independent DINOv2 routes score every retained extent: a reciprocal support–region route and a dense positive-versus-negative localisation route. Third, the strongest boundary-group hypothesis and strongest identity-scored region are compared. Agreement certifies the closed-group hypothesis; strong disagreement invokes a geometric intersection rule. Thus the final decision operates on already bounded hypotheses rather than by unrestricted dense relabelling.

4.6 Support boundary evidence

For each frozen encoder, UBET extracts foreground tokens $F_s^+ = \{\phi(x_s)_u : m_s(u) = 1\}$. It also extracts an exterior-ring bank $F_s^- = \{\phi(x_s)_u : u \in \mathrm{ring}(m_s)\}$. Each bank is deterministically limited to 256 tokens. Their normalised means define foreground and local-background prototypes. Query tokens yield a foreground-versus-background field

$$d(v) = \langle \phi(q)_v, \bar{F}_s^+ \rangle - \langle \phi(q)_v, \bar{F}_s^- \rangle,$$

which is combined with maximum support-set affinity to form a dense semantic field. Its upper 15% supplies support-conditioned query seeds. The same foreground bank is retained for reverse set matching: each support token must find a compatible token inside a candidate. This reciprocal direction discourages a region from being accepted because only one discriminative fragment matches the support.

For the analytic saliency route, the support foreground and its 15-pixel exterior ring define an identity-contrast direction. Foreground and ring tokens are independently sub-

sampled to at most 512 examples per set. Let μ_s^+ and μ_s^- denote the means of the ℓ_2-normalised foreground and ring tokens. We compute

$$w_s = \frac{\mu_s^+ - \mu_s^-}{\|\mu_s^+ - \mu_s^-\|_2}, \qquad h(v) = \sigma\big(w_s^\top \bar{\phi}(q)_v\big)\,,$$

where $\bar{\phi}(q)_v$ is the normalised query token at location v. Min–max normalisation of h produces the support-conditioned saliency field used for seed placement and candidate scoring. The construction is deterministic and closed form. It performs no gradient update, parameter fitting or query-label access, and all pretrained networks remain frozen.

4.7 Boundary-defined query hypotheses

Let $\mathcal{A}_q$ be the deduplicated SAM2 automatic masks. Each binary mask already defines a bounded region. A second route places positive prompts at spatially distinct maxima of the support-conditioned semantic field and obtains one SAM2 mask per prompt. These masks form a directed graph: an edge from mask i to mask j is present when mask i covers the semantic seed associated with j. Weakly connected components define groupable object hypotheses. Within each component, masks are ordered by their mean foreground-minus-ring response and admitted while their uncovered pixels contain more positive than negative evidence. The union of all admitted masks gives the principal closed-group hypothesis g_q; every non-empty component union is retained in $\mathcal{G}_q$. The final candidate set is

$$\mathcal{R}_q = \mathrm{unique}(\mathcal{A}_q \cup \{g_q\} \cup \mathcal{G}_q)\,,$$

where $\mathcal{G}_q$ contains accepted component unions. Candidate generation uses no query label. Exact binary duplicates are removed before scoring.

Let V denote the query token grid, Q_v its normalised DINOv2 token and S^+ the support-foreground bank. For candidate $r \subseteq V$, a 5×5 elliptical kernel defines the inner contour $C(r) = r \setminus \mathrm{erode}(r)$ and exterior ring $E(r) = \mathrm{dilate}(r) \setminus r$. The reciprocal identity

score and boundary contrast are

$$a_{\mathrm{rev}}(r) = \frac{1}{|S^+|} \sum_{s \in S^+} \max_{v \in r} \langle s, Q_v \rangle, \qquad a_{\mathrm{bnd}}(r) = \frac{1}{|r|} \sum_{v \in r} d(v) - \frac{1}{|E(r)|} \sum_{v \in E(r)} d(v).$$

When $E(r)$ is empty, its second term is set to zero. Let $\ell(v)$ be the normalised positive localisation map and $m(v)$ its foreground-minus-background margin. Localisation precision and positive-evidence coverage are

$$a_{\mathrm{loc}}(r) = \frac{1}{|r|} \sum_{v \in r} \ell(v), \qquad a_{\mathrm{pos}}(r) = \frac{\sum_{v \in r} [m(v)]_+}{\sum_{v \in V} [m(v)]_+}.$$

The coverage-and-closure score a_{cov} combines crop-level identity, semantic mass, positive-seed precision and coverage, contour matching on $C(r)$, and foreground purity. Each primitive is computed directly on the frozen token grid. The complete component definitions and fixed coefficients are listed in Supplementary Methods. For N candidates, $\mathcal{R}(x_i) = \mathrm{rank}(x_i)/(N-1)$, with one assigned when $N = 1$. Ranks are clipped below at 0.03 and combined geometrically:

$$s_{\mathrm{pri}}(r) = \Big[\mathcal{R}(a_{\mathrm{rev}})\, \mathcal{R}(a_{\mathrm{cov}})\, \mathcal{R}(a_{\mathrm{bnd}})\Big]^{1/3},$$

$$s_{\mathrm{loc}}(r) = \Big[\mathcal{R}(a_{\mathrm{loc}})\, \mathcal{R}(a_{\mathrm{pos}})\, \mathcal{R}(a_{\mathrm{bnd}})\Big]^{1/3}, \qquad s(r) = \sqrt{s_{\mathrm{pri}}(r) s_{\mathrm{loc}}(r)}.$$

In these expressions juxtaposition denotes multiplication.

4.8 Boundary-conditioned region verification

Let $r_s = \arg\max_{r \in \mathcal{R}_q} s(r)$ be the strongest identity-scored region and let g_q be the support-conditioned closed-group hypothesis. Their agreement is

$$\kappa_q = \mathrm{IoU}(g_q, r_s).$$

The locked conflict certificate uses the fixed threshold $\tau = 0.25$:

$$\hat{m}_q = \begin{cases} g_q, & \kappa_q \geq \tau, \\ g_q \cap r_s, & \kappa_q < \tau. \end{cases}$$

The first branch preserves the complete boundary-defined object when structural and identity evidence are compatible. The second removes unsupported parts when the independently selected identity region strongly disagrees with the boundary group. The query mask is not used to choose the branch or threshold. Diagnostics additionally report the boundary-only g_q, semantic-only r_s and unconditional-intersection alternatives on the identical frozen candidate pool.

4.9 Boundary-source independence

Every independently implemented system constructs its own population of closed regions. Maps from DiffusionEdge and Depth Anything V2 Small are converted into candidate regions; SLIC, Felzenszwalb and Canny provide conventional region or edge partitions. Each system then applies frozen DINOv2 support evidence, provider-conditioned closure and a common region-level arbitration template. Together, these systems test the portability of the inference order rather than define task-specific architectures. Supplementary Methods 1 specifies provider preprocessing, closed-hypothesis construction, support-conditioned evidence, candidate arbitration and the fixed configuration of each independent implementation.

4.10 UBET decision and backbone ablations

All decision variants reuse the candidate masks and evidence computed by one frozen inference pass. Boundary-only returns the closed-group hypothesis g_q; semantic-only returns r_s; fixed intersection always returns $g_q \cap r_s$; and the full certificate uses the rule in Section 4.8. The matched difference between the full and boundary-only paths is calculated per query. The no-boundary-contrast control removes a_{bnd} from both geometric

score products and recomputes the selected region and conflict decision on the unchanged candidate pool; all other primary and localisation signals are retained. Backbone controls replace the frozen DINOv2 feature configuration. ViT-B/14, ViT-L/14 and their concatenated B+L representation use identical image sizes, SAM2 hyperparameters, score definitions and conflict threshold. Prompt-derived candidates are regenerated because prompt locations are themselves obtained from the selected DINOv2 features.

4.11 Hypothesis oracle and boundary diagnostics

For each diagnostic query, the oracle is $\max_{h \in \mathcal{H}_q} \mathrm{IoU}(h, m_q)$, computed only after the prediction has been fixed. Operationally, this is the larger of the best retained-candidate IoU and the IoU of the certificate output. Selection regret is the oracle IoU minus the IoU of the full-certificate output after both masks have been mapped to the benchmark evaluation resolution. Binary morphological boundaries are extracted from prediction and ground truth. Boundary precision and boundary recall match contour pixels within two working-grid pixels; their harmonic mean is reported as the boundary F-score (BF score). This contour diagnostic is distinct from Dice or a region-level classification F1 score, neither of which is reported. Empty predictions remain in all summaries.

For the evidence-ranking audit, every retained closed candidate and its fused deployable score were saved for all 2,400 FSS-1000 queries in the complete 240-class test set. Candidate IoU was computed only after inference. Spearman rank correlation was first measured within each query. The reported pooled coefficient centres candidate ranks within query before combining them. Score-ordering curves group candidates by within-query score percentile. Percentile-bin confidence intervals resample queries as clusters. Top-k coverage is the fraction of queries for which the highest-IoU candidate among the top-k evidence-ranked hypotheses lies within one IoU point of the full candidate oracle. Raw scores are not compared across queries and are not interpreted as probabilities.

4.12 Comparator provenance and protocol harmonisation

All comparator rows use released model code together with the public pretrained checkpoints specified or required by each implementation. Native benchmark reproduction and official-code cross-benchmark evaluation are kept distinct. GF-SAM exposes native PASCAL-5^i and COCO-20^i launchers. Matcher and INSID3 expose native COCO-20^i evaluation and use an adapter for PASCAL-5^i. FSS-SAM3 uses its released visual-only pairwise inference under the common episode manifests. PerSAM, PerSAM-F and ABCDFSS use adapters for both four-fold benchmarks. The remaining released-code evaluations follow the same rule: an adapter may read a locked episode manifest, restore coordinates and accumulate foreground IoU, but it may not change model weights, prompts, score functions or selection parameters. Thus, "released-code" does not imply that every benchmark launcher was published by the original authors.

4.13 Statistical analysis and reproducibility

Runtime and candidate-count diagnostics make the computational scope explicit. Runtime is wall-clock query inference after one-time model initialisation and includes proposal generation, feature extraction and candidate selection. The common-hardware comparison uses one RTX 3090, serial GPU execution and no concurrent inference process (Supplementary Table 8). Total parameter counts are sums of tensors in the exact released checkpoints; the INSID3 count is obtained by instantiating its official DINOv3-L/16 architecture. Because the released repositories expose different profiling interfaces, the timing table is a practical common-hardware profile rather than an architecture-intrinsic speed ranking. Episode-specific variables are disclosed separately and are not used to hide the pretrained footprint. Candidate count is the mean number of unique closed hypotheses retained before final identity selection. A separate three-episode COCO-20^i probe measured 4,214 MiB peak allocated GPU memory for the locked four-by-four configuration; it is used only for resource profiling and not for accuracy or latency estimation.

Let N be the retained candidate count, K_s the number of support tokens, K_r the number of tokens in one candidate and $H \times W$ the localisation grid. Reciprocal support–

candidate matching costs $O(NK_sK_r)$, localisation accumulation costs $O(NHW)$, and query-local ranking costs $O(N \log N)$. Candidate masks and token similarities are scored in bounded batches and released after aggregation. Memory use is therefore controlled by the candidate batch size, while runtime grows approximately linearly with N apart from the final rank operation.

Scores use the aggregation units specified in the dataset protocol and the headline tables, and no query is removed as an outlier. The study reports descriptive effect sizes for controlled comparisons and does not perform a family of null-hypothesis tests.

The UBET inference source used for headline evaluation is version-locked. Diagnostic runners expose candidate masks and evidence vectors without changing headline predictions. Released comparator algorithms and checkpoints are left unchanged; compatibility adapters are restricted to data ingestion, coordinate restoration and metric reporting. The inference and evaluation code is publicly available through the GitHub repository specified below.

4.14 Use of generative artificial intelligence

A large language model was used to assist code development, English-language editing, figure layout, structural consistency checks and LaTeX quality assurance. It did not generate synthetic experimental observations or alter measured outputs. All reported values and image-grounded panels were verified against the underlying result files and source images, and the authors retained responsibility for experimental design, result selection, interpretation and the scientific conclusions.

Data availability

The study uses the publicly available FSS-1000, PASCAL-5^i, COCO-20^i, PerSeg, DAVIS, ISIC 2017, Kvasir-SEG and ETIS datasets. Dataset sources, evaluation partitions, fold definitions and support–query construction are described in the Methods and the corresponding dataset references. The original datasets remain subject to their respective

licences, and no new dataset was generated in this study.

Code availability

The inference and evaluation code is available in the public GitHub repository.

References

[1] 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.

[2] 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)*, pages 4015–4026, 2023. doi: 10.1109/ICCV51070.2023.00371. URL https://doi.org/10.1109/ICCV51070.2023.00371.

[3] 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.

[4] 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.

[5] 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*, 2024. URL https://openreview.net/forum?id=6Gzkhoc6YS.

[6] 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.

[7] 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)*, pages 9196–9205. IEEE, 2019. doi: 10.1109/ICCV.2019.00929. URL https://doi.org/10.1109/ICCV.2019.00929.

[8] 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)*, pages 6941–6952, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Min_Hypercorrelation_Squeeze_for_Few-Shot_Segmentation_ICCV_2021_paper.html.

[9] 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/.

[10] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical Dense Correlation Distillation for Few-Shot Segmenta-

tion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23641–23651, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.02264.

[11] 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.

[12] 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.

[13] 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/.

[14] Suho Park, Subeen Lee, Sangeek Hyun, Hyun Seok Seong, and Jae-Pil Heo. Task-disruptive background suppression for few-shot segmentation. In *AAAI Conference on Artificial Intelligence*, 2023.

[15] 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)*, pages 3380–3390, 2024.

[16] Jintao Tong, Yixiong Zou, Guangyao Chen, Yuhua Li, and Ruixuan Li. Self-Disentanglement and Re-Composition for Cross-Domain Few-Shot Segmentation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=LkgwdSzlb6.

[17] Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment any-

thing semantically. In *Advances in Neural Information Processing Systems*, volume 37, pages 33232–33261, 2024. doi: 10.52202/079017-1047. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3a2ef31a1e45908901adc0ca853a8faf-Abstract-Conference.html.

[18] 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*, pages 21659–21668, 2026. URL https://arxiv.org/abs/2604.12113.

[19] 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.

[20] 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.

[21] 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.

[22] 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)*, pages 1130–1140, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_SegGPT_Towards_Segmenting_Everything_in_Context_ICCV_2023_paper.html.

[23] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and

Yong Jae Lee. Segment Everything Everywhere All at Once. *ArXiv*, abs/2304.06718, 2023. URL https://arxiv.org/abs/2304.06718.

[24] Feng Li, Hao Zhang, Pei Sun, Xueyan Zou, Siyi Liu, Jianwei Yang, Chun-yue Li, Lei Zhang, and Jianfeng Gao. Semantic-SAM: Segment and Recognize Anything at Any Granularity. *ArXiv*, abs/2307.04767, 2023. URL https://arxiv.org/abs/2307.04767.

[25] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499, 2023. URL https://arxiv.org/abs/2303.05499.

[26] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10995–11005, 2022.

[27] Monika Wysocza'nska, Oriane Siméoni, Michael Ramamonjisoa, Andrei Bursuc, Tomasz Trzci'nski, and Patrick P'erez. Clip-dinoiser: Teaching clip a few dino tricks. In *European Conference on Computer Vision*, 2023.

[28] Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, Yansong Tang, Jie Zhou, and Jiwen Lu. Self-Calibrated CLIP for Training-Free Open-Vocabulary Segmentation. *IEEE Transactions on Image Processing*, 34:8271–8284, 2025. doi: 10.1109/TIP.2025.3639996. URL https://doi.org/10.1109/TIP.2025.3639996.

[29] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_

Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

[30] 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.

[31] 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 Doll'ar, 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.

[32] Tianrun Chen, Ankang Lu, Lanyun Zhu, Chao Ding, Chunan Yu, Deyi Ji, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. *ArXiv*, abs/2408.04579, 2024. URL https://arxiv.org/abs/2408.04579.

[33] Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. Sam2-unet: segment anything 2 makes strong

encoder for natural and medical image segmentation. *Visual Intelligence*, 4, 2024. URL https://arxiv.org/abs/2408.08870.

[34] Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. Samwise: Infusing wisdom in sam2 for text-driven video segmentation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3395–3405, 2024.

[35] Johan Wagemans, James H. Elder, Michael Kubovy, Stephen E. Palmer, Mary A. Peterson, Manish Singh, and Ruediger von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological Bulletin*, 138(6):1172–1217, 2012. doi: 10.1037/a0029333.

[36] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurélien Lucchi, Pascal V. Fua, and Sabine Süsstrunk. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34: 2274–2282, 2012. doi: 10.1109/TPAMI.2012.120. URL https://doi.org/10.1109/TPAMI.2012.120.

[37] Leo J. Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1768–1783, 2006.

[38] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Scholkopf. Learning with local and global consistency. In *Neural Information Processing Systems*, 2003.

[39] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Neural Information Processing Systems*, 2013.

[40] 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)*, pages 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.

[41] Yunfan Ye, Kai Xu, Yuhang Huang, Renjiao Yi, and Zhiping Cai. DiffusionEdge: Diffusion probabilistic model for crisp edge detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6675–6683, 2024. doi: 10.1609/aaai.v38i7.28490.

[42] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, volume 37, pages 21875–21911, 2024. doi: 10.52202/079017-0688. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/26cfdcd8fe6fd75cc53e92963a656c58-Abstract-Conference.html.

[43] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986. doi: 10.1109/TPAMI.1986.4767851.

[44] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59:167–181, 2004. doi: 10.1023/B:VISI.0000022288.19776.77.

[45] 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)*, pages 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.

[46] 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.

[47] 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)*, pages 168–172. IEEE, 2018. doi: 10.1109/ISBI.2018.8363547. URL https://doi.org/10.1109/ISBI.2018.8363547.

[48] 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*, pages 451–462. Springer International Publishing, 2019. doi: 10.1007/978-3-030-37734-2_37. URL https://doi.org/10.1007/978-3-030-37734-2_37.

[49] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, 9:283–293, 2014. doi: 10.1007/s11548-013-0926-3.

Author contributions

E.Z. conceived and designed the study, developed the method and software, conducted the experiments, analysed the results, prepared the figures and wrote the manuscript.

Competing interests

The author declares no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.pdf