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.

Boundary-First Segmentation with Closed-Region Hypotheses

EnBao Zhang

enbaozhang52@gmail.com

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

Research Article

Keywords: Few-shot segmentation, one-shot segmentation, training-free inference, boundary representation, foundation models

Posted Date: September 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10886301/v1

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

Additional Declarations: The authors declare no competing interests.

Boundary-First Segmentation with Closed-Region Hypotheses

Anonymous Author(s)

Abstract

Few-shot segmentation usually infers object extent from dense support–query similarity, leaving spatial coherence to emerge after semantic decisions. We instead formulate Boundary-First one-shot segmentation: closed query regions are constructed before support-conditioned identity arbitration. Unified Boundary Evidence Transport (UBET) implements this order without task-specific training by combining region hypotheses with self-supervised interior, boundary and exterior evidence. The formulation is not tied to a particular proposal model. We instantiate it with SAM2, visual-only SAM3, diffusion and depth boundaries, superpixels, Canny edges and Felzenszwalb regions. Across the full evaluation, UBET is assessed on natural-image, video and medical segmentation, while the alternative boundary providers test the portability of the inference order. Under matched protocols, UBET (SAM2) obtains 73.34% foreground mean intersection-over-union (mIoU) on PASCAL-5^i, 59.34% on COCO-20^i and 87.24% on FSS-1000; UBET (SAM3) reaches 83.19% on PASCAL-5^i. Evaluation across seven benchmarks, matched decision-order interventions and candidate-oracle analyses supports Boundary-First inference as an extensible training-free paradigm for spatially coherent one-shot segmentation. Code is available at the GitHub.

Keywords: Few-shot segmentation, one-shot segmentation, training-free inference, boundary representation, foundation models.

1 Introduction

Image segmentation is commonly formulated as dense labeling: a model assigns a class to each pixel, and an object mask emerges from these local decisions. Few-shot segmentation extends this formulation to unseen categories by using one or a few annotated support examples [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]. Personalized 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.

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 organization, 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 fixed 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 defines a family of segmentation systems. 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 principal instantiation studied here, and six alternative-source variants test the same ordering under substantially different boundary constructions.

The contributions are threefold:

- We formulate one-shot segmentation over a finite family of closed-region hypotheses and derive a stage-wise decomposition that distinguishes candidate availability from candidate-selection error.
- We introduce UBET, a training-free implementation that combines frozen region proposals with support-conditioned interior, boundary and exterior evidence. Its conflict certificate reconciles geometric and identity hypotheses without pixel-wise model adaptation.
- We instantiate the same Boundary-First order with SAM2, visual-only SAM3 and six independent boundary-source systems based on diffusion, depth, superpixels and conventional edges. Their consistent effectiveness indicates that the principle extends beyond a single proposal model. Evaluation on seven benchmarks further covers four-fold evaluation, cross-dataset transfer, decision-order interventions, a 2,400-query candidate audit and resource cost.

2 Related Work

2.1 Few-Shot and Training-Free Segmentation

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]. Several TPAMI studies refine this semantic-to-mask pipeline from complementary directions. PFENet++ improves contextual prior masks and suppresses noisy correspondence responses, HPA develops holistic prototype activation, BAM separates base-class rejection from meta-class prediction, and PAT introduces class-aware prompt transfer [17–20]. These methods improve transferred semantic evidence. Boundary-First instead changes the region hypotheses on which such evidence acts.

Recent training-free systems sharpen semantic prompting in different ways. GF-SAM organizes support–query correspondences as a graph, PR-MaGIC refines prompts through

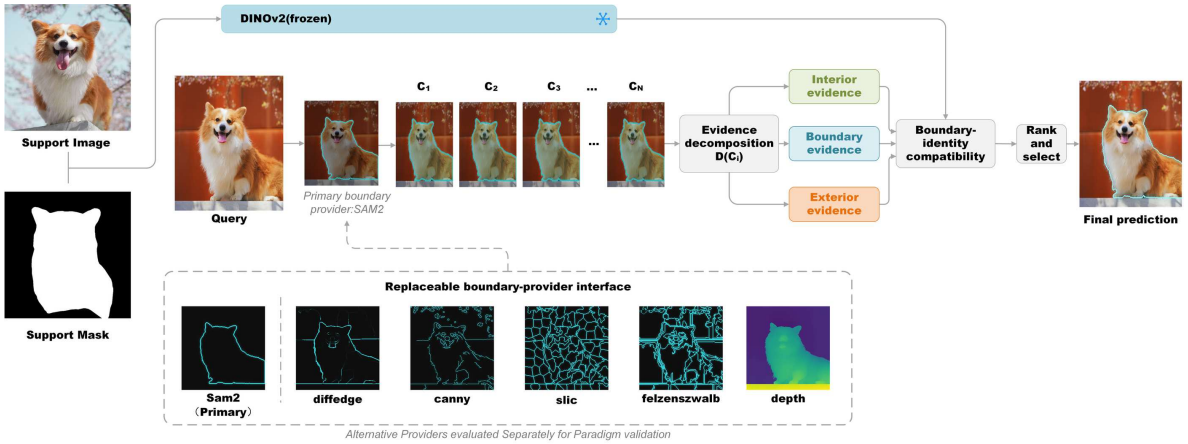


Figure 1: Unified Boundary Evidence Transport (UBET) implements Boundary-First one-shot segmentation as a continuous inference pipeline. An annotated support image defines object identity, while an unlabeled query is first converted into a population of closed region hypotheses. Each hypothesis is decomposed 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.

mask-decoder gradients, and INSID3 and FSSDINO exploit the semantic selectivity of DINOv3 features [21–24]. FSS-SAM3 instead places support and query images on a shared canvas and uses frozen SAM3 for visual-only or concept-prompted transfer [25]. 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.

2.2 Foundation Models and Boundary Representations

The broader foundation-model literature has explored in-context segmentation, promptable recognition, open-world grounding and open-vocabulary dense prediction [26–32]. Self-supervised visual transformers provide the correspondence features used by many of these systems [4, 33, 34], while SAM2 and SAM3 extend promptable segmentation to videos and concepts [3, 35]. Adaptation studies show that these foundation models can be repurposed for downstream segmentation [36–38]. 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 organizing 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 [39–42]. Entropy-regularized optimal transport provides a principled correspondence operator for structured matching [43].

The proposed formulation connects these two lines of work by using explicit region closure to define admissible outputs and support-conditioned matching to select among them. Its contribution is therefore the ordering and unit of inference, not a replacement feature encoder or a newly trained segmentation decoder.

3 Method

3.1 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 unlabeled 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 factorization

$$\begin{aligned} p(h, z = 1 \mid q, \mathcal{S}) = p_\mathcal{B}(h \mid q, \mathcal{S}) \\ \cdot p_\mathcal{I}(z = 1 \mid h, q, \mathcal{S}), \quad h \in \mathcal{H}_\mathcal{B}(q), \end{aligned} \tag{1}$$

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

$$\mathcal{R}_q = \mathcal{B}(q; \mathcal{S}), \qquad g_q = \Gamma(\mathcal{R}_q; \mathcal{S}, q), \tag{2}$$

$$r_s = \arg\max_{r \in \mathcal{R}_q} \Psi(r; \mathcal{S}, q), \qquad \hat{m}_q = \mathcal{C}(g_q, r_s). \tag{3}$$

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.

3.2 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

$$\mathcal{H}_q = \mathcal{R}_q \cup \{\mathcal{C}(g_q, r_s)\}. \tag{4}$$

Define the attainable ceiling and selection regret as

$$O_q = \max_{h \in \mathcal{H}_q} \mathrm{IoU}(h, m_q), \qquad \Delta_q = O_q - \mathrm{IoU}(\hat{m}_q, m_q). \tag{5}$$

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

$$1 - \mathrm{IoU}(\hat{m}_q, m_q) = \underbrace{1 - O_q}_{\text{candidate-generation error}} + \underbrace{\Delta_q}_{\text{selection regret}}. \tag{6}$$

Defining $\epsilon_\mathcal{B} = \mathbb{E}_q[1 - O_q]$ and $\epsilon_\mathcal{I} = \mathbb{E}_q[\Delta_q]$ gives the population relation

$$1 - \mathbb{E}_q[\mathrm{IoU}(\hat{m}_q, m_q)] = \epsilon_\mathcal{B} + \epsilon_\mathcal{I}. \tag{7}$$

The two terms isolate candidate-generation and arbitration errors. A low candidate 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,

$$P_\Psi(G(\mathcal{R}_q)) \neq G(P_\Psi(\mathcal{R}_q)) \,. \tag{8}$$

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.

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

$$\mathrm{IoU}(g \cap r, m) \geq \mathrm{IoU}(g, m) \quad \Longleftrightarrow \quad ax \geq y(A + b). \tag{9}$$

Intersection is therefore beneficial when the identity region removes enough unsupported area (x) to compensate for discarded target area (y). UBET uses it only after detecting conflict.

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 neighborhood of the true contour. Under the regularity conditions for a non-self-overlapping tubular neighborhood, the symmetric-difference area obeys the conservative bound $\delta_d \leq 2dP + \pi d^2$. Whenever $\delta_d < A$,

$$\mathrm{IoU}(r, m) \geq \frac{A - \delta_d}{A + \delta_d}. \tag{10}$$

The bound shows that identical contour displacement produces a larger overlap penalty for targets with higher perimeter-to-area ratio P/A.

The evidence score is a query-local ordinal surrogate rather than a calibrated IoU estimate. Candidate-level rank correlation and top-k oracle coverage test whether it orders hypotheses correctly within each query.

3.3 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 localization 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 relabeling.

3.4 Frozen Visual Evidence Providers

The evaluated UBET implementation combines SAM2 Hiera-L with frozen DINOv2 ViT-B/14 and ViT-L/14 encoders. SAM2 experiments use the Hiera-L checkpoint `sam2.1_hiera_large`; the model is 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 normalized within each encoder; B/L evidence is combined in the reciprocal region-matching and dense-localization 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 optimization or checkpoint adaptation.

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.

3.5 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 normalized means define foreground and local-background prototypes. Query tokens yield a foreground-versus-background field

$$d(v) = \langle \phi(q)_v, \bar{F}_s^+ \rangle - \langle \phi(q)_v, \bar{F}_s^- \rangle, \tag{11}$$

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

$$w_s = \frac{\mu_s^+ - \mu_s^-}{\|\mu_s^+ - \mu_s^-\|_2}, \qquad h(v) = \sigma\Big(w_s^\top \bar{\phi}(q)_v\Big), \tag{12}$$

where $\bar{\phi}(q)_v$ is the normalized query token at location v. Min–max normalization of h produces the support-conditioned saliency field used for seed placement and candidate scoring. This construction is deterministic and closed form.

3.6 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

$$\mathcal{R}_q = \mathrm{unique}(\mathcal{A}_q \cup \{g_q\} \cup \mathcal{G}_q), \tag{13}$$

where $\mathcal{G}_q$ contains accepted component unions. Exact binary duplicates are removed before scoring.

Let V denote the query token grid, Q_v its normalized 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

$$
\begin{aligned}
a_{\mathrm{rev}}(r) &= \frac{1}{|S^+|} \sum_{s \in S^+} \max_{v \in r} \langle s, Q_v \rangle, \\
a_{\mathrm{bnd}}(r) &= \frac{1}{|r|} \sum_{v \in r} d(v) - \frac{1}{|E(r)|} \sum_{v \in E(r)} d(v).
\end{aligned}
\tag{14}
$$

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

$$
a_{\mathrm{loc}}(r) = \frac{1}{|r|} \sum_{v \in r} \ell(v), \qquad a_{\mathrm{pos}}(r) = \frac{\sum_{v \in r} [m(v)]_+}{\sum_{v \in V} [m(v)]_+}. \tag{15}
$$

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 Supplemental Methods. For N candidates, let $\mathrm{rank}(x_i)$ be the zero-based ascending rank, so the largest value has rank $N-1$. We set $\mathcal{R}(x_i) = \mathrm{rank}(x_i)/(N-1)$, with one assigned when $N = 1$. Ranks are clipped below at 0.03 and combined geometrically:

$$
s_{\mathrm{pri}}(r) = \Big[\mathcal{R}(a_{\mathrm{rev}})\, \mathcal{R}(a_{\mathrm{cov}})\, \mathcal{R}(a_{\mathrm{bnd}}) \Big]^{1/3}, \tag{16}
$$

$$
\begin{aligned}
s_{\mathrm{loc}}(r) &= \Big[\mathcal{R}(a_{\mathrm{loc}})\, \mathcal{R}(a_{\mathrm{pos}})\, \mathcal{R}(a_{\mathrm{bnd}}) \Big]^{1/3}, \\
s(r) &= \sqrt{s_{\mathrm{pri}}(r) s_{\mathrm{loc}}(r)}.
\end{aligned}
\tag{17}
$$

In these expressions juxtaposition denotes multiplication.

3.7 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

$$
\kappa_q = \mathrm{IoU}(g_q, r_s). \tag{18}
$$

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

$$
\hat{m}_q = \begin{cases} g_q, & \kappa_q \geq \tau, \\ g_q \cap r_s, & \kappa_q < \tau. \end{cases} \tag{19}
$$

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. Diagnostics additionally report the boundary-only g_q, semantic-only r_s and unconditional-intersection alternatives on the identical frozen candidate pool.

3.8 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 across different boundary representations. Supplemental Methods 1 specifies provider preprocessing, closed-hypothesis construction, support-conditioned evidence, candidate arbitration and the fixed configuration of each independent implementation.

4 Experiments

4.1 Implementation Details 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 [1,7,44]; PerSeg evaluates personalized objects [5]; DAVIS evaluates annotated video objects [45]; and ISIC, Kvasir-SEG and ETIS cover skin lesions and polyps [46–48]. PASCAL-5^i and COCO-20^i use seed-zero four-fold manifests with 1,000 episodes per fold. FSS-1000 uses the complete 240-class, 2,400-image test set. 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 predictions contribute zero IoU.

For natural-image cross-dataset transfer, we first identify semantic categories present in both COCO and PASCAL. Each episode then pairs a support example from the source dataset with a target query of the same shared category. COCO-to- PASCAL and PASCAL-to-COCO each contain 100 target queries. Kvasir-SEG-to-ETIS uses all 182 ETIS queries under the same one-shot protocol.

The evaluation metric is foreground mean intersection-over-union (mIoU). FSS-1000, PASCAL-5^i and COCO-20^i use class aggregation; PerSeg and DAVIS aggregate over object groups or sequences; and the medical benchmarks use query means. Region construction and evidence scoring operate on a 256×256 grid. DINOv2 inputs use 518×518 pixels, while SAM2 retains its native preprocessing. Predictions on PASCAL-5^i and COCO-20^i are restored through the stored query transform before evaluation.

Comparators use their released code and prescribed public checkpoints. GF-SAM provides native PASCAL-5^i and COCO-20^i launchers; Matcher and INSID3 use native COCO-20^i evaluation and a PASCAL-5^i data adapter; FSS-SAM3 uses released visual-only pairwise inference; and PerSAM, PerSAM-F and ABCDFSS use data adapters for both four-fold benchmarks. These adapters load the episode, restore coordinates and accumulate foreground IoU while retaining each released model's weights, prompts, scores and selection parameters.

Runtime is wall-clock query inference after one-time initialization and includes proposal generation, feature extraction and candidate selection. All methods are profiled serially on one RTX 3090 (Table S7). Parameter counts include every instantiated pretrained checkpoint; INSID3 uses its official DINOv3-L/16 architecture. The evaluated UBET configuration retains 39.47 unique hypotheses per query on average and allocates 4,214 MiB at peak in a three-episode COCO-20^i probe. For N candidates, K_s support tokens, K_r candidate tokens and an $H \times W$ localization grid, reciprocal matching costs $O(NK_sK_r)$, localization costs $O(NHW)$ and ranking costs $O(N \log N)$.

4.2 One-Shot Segmentation Across Seven Benchmarks

Tables 1 and 2 report the matched one-shot comparisons. SAM2 and DINOv2 remain frozen, and support-conditioned saliency is computed analytically from foreground and exterior-ring prototypes. Main results and mIoU ablations use each benchmark's evaluation resolution. Mechanistic diagnostics use UBET (SAM2) unless another 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

Table 1: Four-Fold One-Shot 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. 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
PerSAM [5]	50.18	53.35	45.27	39.98	47.20
PerSAM-F [5]	46.20	51.03	46.69	38.31	45.56
FSSDINO [24]	51.46	62.56	58.82	51.46	56.08
PR-MaGIC [22]	66.65	67.36	63.17	65.54	65.68
INSID3 [23]	63.77	73.14	67.96	68.42	68.32
Matcher [6]	67.66	70.29	66.05	67.69	67.92
GF-SAM [21]	70.61	75.78	68.32	72.40	71.78
ABCDFSS [49]	35.17	43.83	40.97	36.51	39.12
FSS-SAM3 [25]	<u>76.00</u>	<u>81.14</u>	<u>82.27</u>	<u>84.71</u>	<u>81.03</u>
UBET (SAM2)	74.69	77.02	68.58	73.08	73.34
UBET (SAM3)	**81.15**	**82.93**	**83.08**	**85.61**	**83.19**

b COCO-20^i

Method	Fold 0	Fold 1	Fold 2	Fold 3	Mean
PerSAM [5]	22.62	23.74	21.40	20.79	22.14
PerSAM-F [5]	21.85	23.59	23.13	22.34	22.73
FSSDINO [24]	38.72	41.33	40.12	42.44	40.65
PR-MaGIC [22]	50.46	53.65	56.32	53.13	53.39
INSID3 [23]	52.34	55.41	54.27	53.82	53.96
Matcher [6]	51.48	52.76	51.22	52.72	52.05
GF-SAM [21]	56.56	**60.96**	<u>59.33</u>	<u>57.71</u>	<u>58.64</u>
ABCDFSS [49]	22.02	25.40	24.24	24.35	24.00
FSS-SAM3 [25]	56.65	54.26	53.80	55.32	55.01
UBET (SAM2)	<u>57.64</u>	<u>60.13</u>	**60.11**	**59.46**	**59.34**
UBET (SAM3)	**58.54**	58.29	58.85	55.12	57.70

(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 neighboring 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 measured 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.

4.3 Boundary-First Arbitration Transfers Across Datasets

We next changed the dataset from which the one-shot support was drawn. For COCO-to-PASCAL and PASCAL-to-COCO, episodes are formed only from semantic categories that occur in both datasets: the support comes from the source dataset and the query depicts the same category in the target dataset. Kvasir-SEG-to-ETIS evaluates transfer between independent endoscopic collections.

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 with ETIS queries yields 67.40% (Table 3). Replacing the proposal provider with SAM3 changes these scores

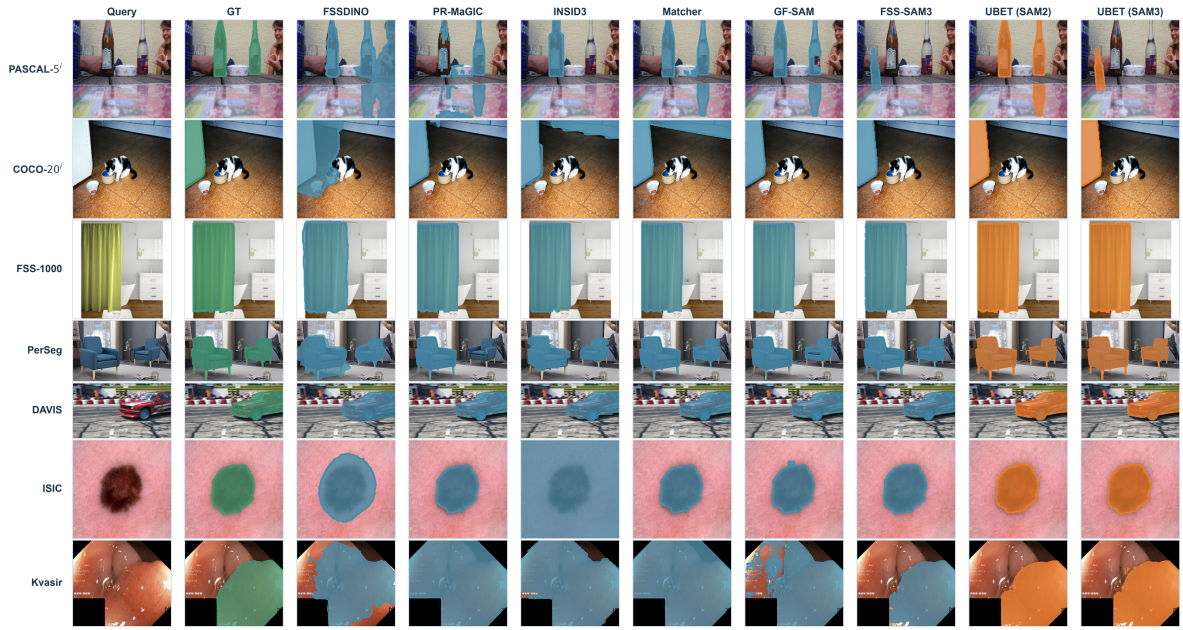


Figure 2: Qualitative comparison across seven one-shot segmentation benchmarks. The query and ground truth are followed by predictions from FSSDINO, PR-MaGIC, INSID3, Matcher, GF-SAM and visual-only FSS-SAM3, followed by the SAM2 and SAM3 UBET instantiations. Every method receives the same support–query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

Table 2: One-Shot Segmentation Across Seven Benchmarks. Values are foreground mIoU percentages on the complete query partitions under the declared protocols. Comparator provenance and adapter status are documented in Section 4.1. 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 [21]	<u>85.5</u>	87.2	70.6	44.2	60.4
INSID3 [23]	83.1	90.1	73.5	52.2	43.9
FSSDINO [24]	75.2	82.1	60.2	48.8	41.4
PR-MaGIC [22]	84.7	94.1	<u>79.6</u>	54.0	56.8
ABCDFSS [49]	69.2	82.8	56.4	44.2	55.1
FSS-SAM3 [25]	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**

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 shows that transfer remains sensitive to the proposal model. Both UBET instantiations nevertheless improve on direct FSS-SAM3 in the two natural-image directions. These experiments measure how each complete inference rule transfers when the support and query originate from different datasets.

Table 3: Cross-Dataset One-Shot Segmentation. Natural-image episodes pair source support and target queries from categories shared by COCO and PASCAL, using 100 target queries in each 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 [24]	41.49	33.80	30.12
PR-MaGIC [22]	35.69	36.48	29.61
INSID3 [23]	46.70	39.30	48.60
Matcher [6]	39.91	42.04	34.30
GF-SAM [21]	<u>50.75</u>	43.91	50.28
ABCDFSS [49]	32.13	26.81	27.81
FSS-SAM3 [25]	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: Boundary-Source Transfer Across Independent Implementations. 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

4.4 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 [40,50–53]. 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. Provider rankings vary across datasets, while every implementation retains useful segmentation under a different candidate construction. Provider-specific details are given in Supplemental Methods 1.

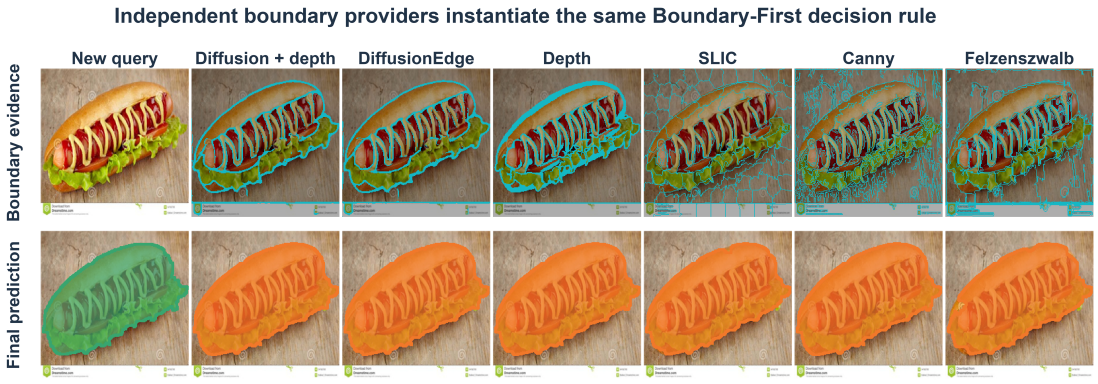


Figure 3: Image-grounded transfer across independent boundary providers. The upper row shows the query and the evidence supplied by fused diffusion–depth cues, DiffusionEdge, depth boundaries, SLIC, Canny and Felzenszwalb. The lower row shows the ground truth and the corresponding final predictions.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

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.

4.5 Ablations of Conflict Repair and Identity Evidence

The 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 are consistent with Proposition 3: conflict-conditioned reconciliation improves the average result even though the query-level improvement criterion need not hold for every episode.

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.

For the diagnostic analysis, the candidate oracle is the highest IoU among the retained hypotheses and the certificate output. Selection regret is the oracle IoU minus the selected-mask IoU. The boundary F-score is the harmonic mean of contour precision and recall under a two-pixel working-grid tolerance.

Additional complete-path controls and frozen DINOv2 feature-scale comparisons are reported in Tables S2 and S3; sensitivity to the shared conflict threshold is reported in Table S4. 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.

Table 5: Mechanism Ablations 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 retrospective 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.

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

On DAVIS and Kvasir-SEG, semantic-only selection and unconditional intersection retain only a discriminative object fragment, whereas the closed boundary hypothesis 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 retrospective 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 (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 indicate 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.

4.6 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

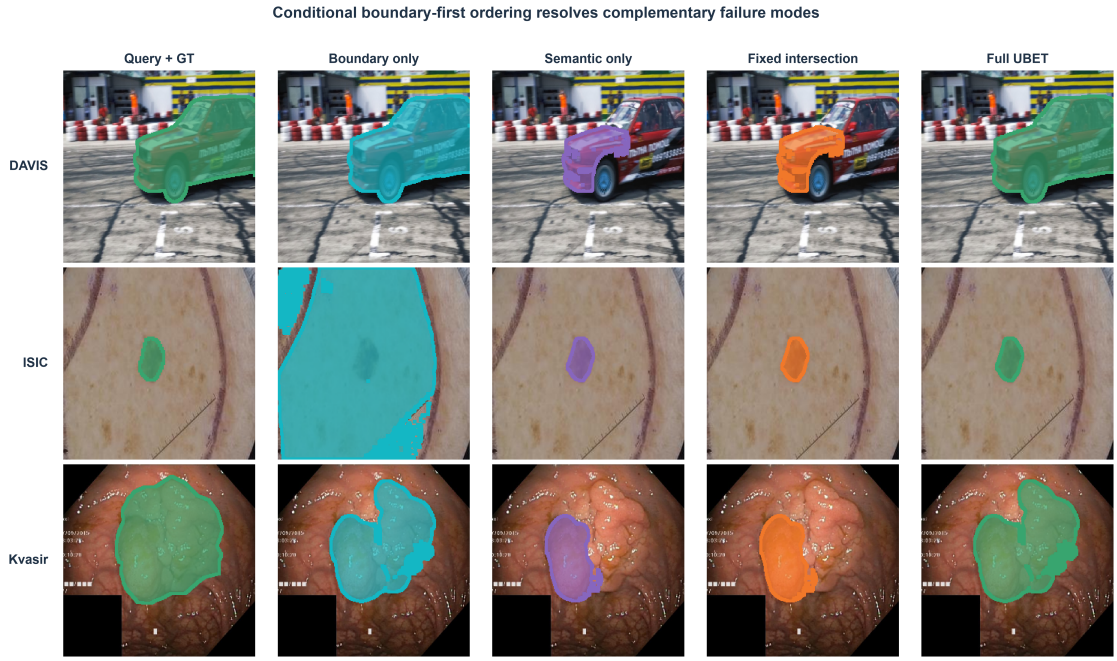


Figure 4: Complementary failure modes of fixed decision paths. Three real episodes compare boundary-only, semantic-only, unconditional-intersection and full conditional inference using identical frozen evidence. Boundary evidence preserves object extent in the DAVIS and Kvasir-SEG examples, whereas identity 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.

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. PerSeg reaches a BF score of 86.97%, DAVIS reaches 81.01%, and ISIC and Kvasir-SEG have substantially weaker contour scores. The lower BF scores on ISIC and Kvasir-SEG are consistent with their more difficult candidate geometry. Their retained high-IoU candidates also show that selection remains a separate source of error.

4.7 Accuracy, Footprint, and Inference Cost

We profiled the released-code comparators under serial execution after one-time model initialization. Total parameters count all instantiated pretrained weights because reporting only trainable parameters would hide the dominant foundation-model footprint. The accuracy column in Table 6 is the descriptive arithmetic mean of the seven dataset-specific results. It summarizes 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 [34]; 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, 75.93% and 75.87%, respectively. The SAM2 model exceeds PR-MaGIC by 6.12 points, and its 615.40

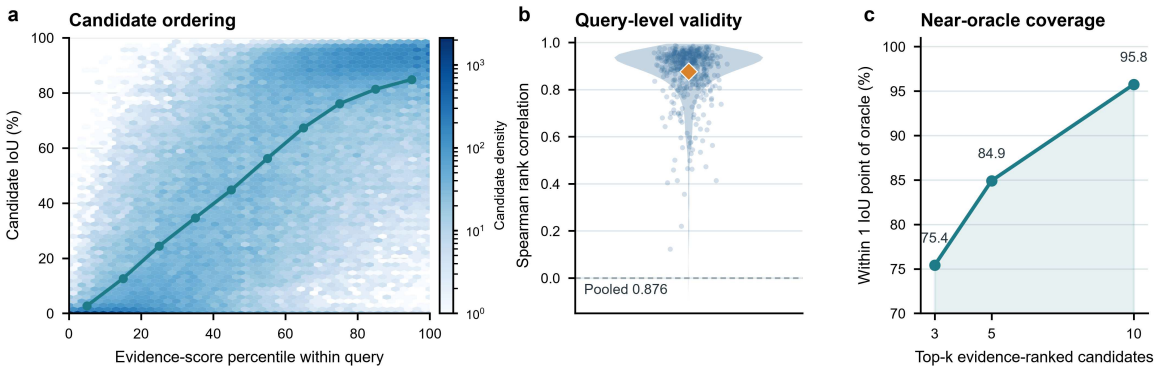


Figure 5: Query-local evidence-score validity across all retained candidates. **a**, Density of candidate IoU across within-query evidence-score percentiles for 74,061 candidates from all 2,400 FSS-1000 queries; the line shows query-aggregated percentile-bin means. **b**, Distribution of query-level Spearman rank correlation coefficients; the orange diamond marks the pooled within-query coefficient. **c**, Near-oracle coverage among the top three, five and ten evidence-ranked hypotheses.

Table 6: Accuracy, Model Footprint, and Inference Cost. Accuracy is the descriptive mean foreground mIoU across seven benchmarks. Parameters include all instantiated pretrained weights; runtime is end-to-end serial query latency after initialization on one RTX 3090. Bold and underlined entries denote the best and second-best values; lower is better for parameters and runtime.

Method	Mean mIoU	Parameters (M)	Time (s/query)
PerSAM [5]	46.13	641.09	0.928
PerSAM-F [5]	54.37	641.09	7.376
FSSDINO [24]	57.78	<u>85.70</u>	**0.138**
PR-MaGIC [22]	69.75	945.46	11.682
INSID3 [23]	66.44	303.15	0.7331
Matcher [6]	64.47	945.46	6.244
GF-SAM [21]	68.33	945.46	0.736
ABCDFSS [49]	52.97	**25.56**	2.821
FSS-SAM3 [25]	71.06	840.51	<u>0.698</u>
UBET (SAM2)	<u>75.87</u>	615.40	2.481
UBET (SAM3)	**75.93**	1,231.46	1.8146

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 (Table 6). 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 (Table S7). Together, these stages account for 77.96% of latency. The final certificate contributes approximately 0.01%. The evaluated 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.

5 Discussion

The central result is that closed regions can serve as the decision units of one-shot segmentation before identity arbitration. This ordering remains effective in independent implementations based on diffusion, depth, superpixels, Canny and Felzenszwalb regions, so the result is not tied to one proposal foundation model. The same UBET implementation is evaluated across category, personalized-object, video and medical benchmarks, with additional cross-dataset support shifts. These experiments jointly assess Boundary-First segmentation as an inference paradigm and UBET as one concrete implementation.

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%. On COCO-20^i, SAM2-based UBET remains stronger, at 59.34% versus 57.70%. The complementary ranking shows that closed-region evidence and arbitration transfer to SAM3 while the preferred proposal provider can remain dataset dependent.

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. The measured controls accordingly favor conditional repair over unconditional intersection.

The decision-order intervention tests more than the presence of two useful signals. As formalized by Proposition 2, closing a contour and pruning its atoms by identity evidence are generally non-commutative operations. Boundary-First processing improves over Identity-First processing in all five tested domains, indicating that retaining the complete geometric hypothesis before identity arbitration matters in these settings.

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 each 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 fused score as a query-local candidate selector on FSS-1000. Extending this calibration across datasets remains an open problem.

Appearance and boundary evidence remain complementary. Single-backbone and alternate-path controls vary in their relative ranking across domains, so we retain one fixed 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 (Tables S5 and S6).

UBET also inherits errors from both frozen evidence providers. Sparse SAM2 sampling can omit thin or weakly bounded targets, while pretrained DINOv2 correspondences can favor a visually related distractor or only the most diagnostic object part. The same conflict threshold is used across datasets. The method is training-free: it performs no dataset-level task-specific

training, episodic gradient optimization 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 favorable 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 main comparison.

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.

6 Conclusion

This work establishes a simple ordering principle for one-shot segmentation: construct a closed object extent before assigning the support identity. UBET operationalizes this principle with frozen SAM and DINO features, while six independently implemented systems based on diffusion, depth, superpixels and conventional edges reproduce effective Boundary-First behavior. Their results, together with the matched order interventions, support the ordering itself rather than a particular proposal model. Boundary-First segmentation therefore opens a broader design space in which learned boundary providers, region-level matching and self-learning arbitration can improve candidate availability and selection independently.

References

[1] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-Shot Learning for Semantic Segmentation," *ArXiv*, vol. abs/1709.03410, 2017. [Online]. Available: https://arxiv.org/abs/1709.03410

[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. B. Girshick, "Segment Anything," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023. [Online]. Available: https://doi.org/10.1109/ICCV51070.2023.00371

[3] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. K. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C. Wu, R. B. Girshick, P. Dollár, and C. Feichtenhofer, "SAM 2: Segment Anything in Images and Videos," in *International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=Ha6RTeWMd0

[4] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. B. Huang, S.-W. Li, I. Misra, M. G. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning Robust Visual

Features without Supervision," *ArXiv*, vol. abs/2304.07193, 2023. [Online]. Available: https://arxiv.org/abs/2304.07193

[5] R. Zhang, Z. Jiang, Z. Guo, S. Yan, J. Pan, H. Dong, Y. Qiao, P. Gao, and H. Li, "Personalize Segment Anything Model with One Shot," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=6Gzkhoc6YS

[6] Y. Liu, M. Zhu, H. Li, H. Chen, X. Wang, and C. Shen, "Matcher: Segment Anything with One Shot Using All-Purpose Feature Matching," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=yzRXdhk2he

[7] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-Shot Image Semantic Segmentation With Prototype Alignment," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 9196–9205. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00929

[8] J. Min, D. Kang, and M. Cho, "Hypercorrelation Squeeze for Few-Shot Segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6941–6952. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Min_Hypercorrelation_Squeeze_for_Few-Shot_Segmentation_ICCV_2021_paper.html

[9] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior Guided Feature Enrichment Network for Few-Shot Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 1050–1065, Feb. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9154595/

[10] B. Peng, Z. Tian, X. Wu, C. Wang, S. Liu, J. Su, and J. Jia, "Hierarchical Dense Correlation Distillation for Few-Shot Segmentation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 23 641–23 651.

[11] W. Liu, C. Zhang, H. Ding, T.-Y. Hung, and G. Lin, "Few-Shot Segmentation With Optimal Transport Matching and Message Flow," *IEEE Transactions on Multimedia*, vol. 25, pp. 5130–5141, 2023.

[12] S. Lei, X. Zhang, J. He, F. Chen, B. Du, and C.-T. Lu, "Cross-Domain Few-Shot Semantic Segmentation," in *European Conference on Computer Vision*, 2022. [Online]. Available: https://doi.org/10.1007/978-3-031-20056-4_5

[13] X. Huang, C. Zhu, and W. Chen, "RestNet: Boosting Cross-Domain Few-Shot Segmentation with Residual Transformation Network," in *British Machine Vision Conference*, 2023. [Online]. Available: https://proceedings.bmvc2023.org/12/

[14] S. Park, S. Lee, S. Hyun, H. S. Seong, and J.-P. Heo, "Task-disruptive background suppression for few-shot segmentation," in *AAAI Conference on Artificial Intelligence*, 2023.

[15] J. Nie, Y. Xing, G. Zhang, P. Yan, A. Xiao, Y.-P. Tan, A. C. Kot, and S. 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.

[16] J. Tong, Y. Zou, G. Chen, Y. Li, and R. Li, "Self-Disentanglement and Re-Composition for Cross-Domain Few-Shot Segmentation," in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=LkgwdSzlb6

[17] X. Luo, Z. Tian, T. Zhang, B. Yu, Y. Y. Tang, and J. Jia, "PFENet++: Boosting few-shot semantic segmentation with the noise-filtered context-aware prior mask," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 2, pp. 1273–1289, 2024.

[18] G. Cheng, C. Lang, and J. Han, "Holistic prototype activation for few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4650–4666, 2023.

[19] C. Lang, G. Cheng, B. Tu, C. Li, and J. Han, "Base and meta: A new perspective on few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 669–10 686, 2023.

[20] H. Bi, Y. Feng, W. Diao, P. Wang, Y. Mao, K. Fu, H. Wang, and X. Sun, "Prompt-and-transfer: Dynamic class-aware enhancement for few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 1, pp. 131–148, 2025.

[21] A. Zhang, G. Gao, J. Jiao, C. H. Liu, and Y. Wei, "Bridge the points: Graph-based few-shot segment anything semantically," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 33 232–33 261. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/hash/3a2ef31a1e45908901adc0ca853a8faf-Abstract-Conference.html

[22] M. Lee, S. Hur, S. Hwang, and W. H. 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*, 2026, pp. 21 659–21 668. [Online]. Available: https://arxiv.org/abs/2604.12113

[23] C. Cuttano, G. Trivigno, C. Reich, D. Cremers, C. Masone, and S. Roth, "Insid3: Training-free in-context segmentation with dinov3," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2026/papers/Cuttano_INSID3_Training-Free_In-Context_Segmentation_with_DINOv3_CVPR_2026_paper.pdf

[24] H. M. Zakir and E. T. W. Ho, "Revealing the semantic selection gap in DINOv3 through training-free few-shot segmentation," *arXiv preprint arXiv:2602.07550*, 2026. [Online]. Available: https://arxiv.org/abs/2602.07550

[25] Y.-J. Tsai, Y.-Y. Lin, and C.-Y. Wang, "Few-shot semantic segmentation meets SAM3," *arXiv preprint arXiv:2604.05433*, 2026. [Online]. Available: https://arxiv.org/abs/2604.05433

[26] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, "SegGPT: Towards Segmenting Everything in Context," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1130–1140. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/Wang_SegGPT_Towards_Segmenting_Everything_in_Context_ICCV_2023_paper.html

[27] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Gao, and Y. J. Lee, "Segment Everything Everywhere All at Once," *ArXiv*, vol. abs/2304.06718, 2023. [Online]. Available: https://arxiv.org/abs/2304.06718

[28] F. Li, H. Zhang, P. Sun, X. Zou, S. Liu, J. Yang, C.-y. Li, L. Zhang, and J. Gao, "Semantic-SAM: Segment and Recognize Anything at Any Granularity," *ArXiv*, vol. abs/2307.04767, 2023. [Online]. Available: https://arxiv.org/abs/2307.04767

[29] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. yue Li, J. Yang, H. Su, J.-J. Zhu, and L. Zhang, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *ArXiv*, vol. abs/2303.05499, 2023. [Online]. Available: https://arxiv.org/abs/2303.05499

[30] X. Dong, J. Bao, Y. Zheng, T. Zhang, D. Chen, H. Yang, M. Zeng, W. Zhang, L. Yuan, D. Chen, F. Wen, and N. Yu, "Maskclip: Masked self-distillation advances contrastive language-image pretraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 995–11 005.

[31] M. Wysoczańska, O. Siméoni, M. Ramamonjisoa, A. Bursuc, T. Trzciński, and P. Pérez, "Clip-dinoiser: Teaching clip a few dino tricks," in *European Conference on Computer Vision*, 2024.

[32] S. Bai, Y. Liu, Y. Han, H. Zhang, Y. Tang, J. Zhou, and J. Lu, "Self-Calibrated CLIP for Training-Free Open-Vocabulary Segmentation," *IEEE Transactions on Image Processing*, vol. 34, pp. 8271–8284, 2025. [Online]. Available: https://doi.org/10.1109/TIP.2025.3639996

[33] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9630–9640. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html

[34] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. E. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski, "DINOv3," *Transactions on Machine Learning Research*, 2026, featured Certification. [Online]. Available: https://openreview.net/forum?id=2NlGyqNjns

[35] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Surís, C. K. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Radle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. Hazra, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H. K. Cheng, P. Doll'ar, N. Ravi, K. Saenko, P. Zhang, and C. Feichtenhofer, "SAM 3: Segment anything with concepts," in *International Conference on Learning Representations*, 2026. [Online]. Available: https://proceedings.iclr.cc/paper_files/paper/2026/hash/e0982cbc81401df3430ee1ff780dc7a2-Abstract-Conference.html

[36] T. Chen, A. Lu, L. Zhu, C. Ding, C. Yu, D. Ji, Z. Li, L. Sun, P. Mao, and Y. Zang, "Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more," *ArXiv*, vol. abs/2408.04579, 2024. [Online]. Available: https://arxiv.org/abs/2408.04579

[37] X. Xiong, Z. Wu, S. Tan, W. Li, F. Tang, Y. Chen, S. Li, J. Ma, and G. Li, "Sam2-unet: segment anything 2 makes strong encoder for natural and medical image segmentation," *Visual Intelligence*, vol. 4, 2024. [Online]. Available: https://arxiv.org/abs/2408.08870

[38] C. Cuttano, G. Trivigno, G. Rosi, C. Masone, and G. Averta, "Samwise: Infusing wisdom in sam2 for text-driven video segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 3395–3405.

[39] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh, and R. von der Heydt, "A century of gestalt psychology in visual perception: I. perceptual

grouping and figure–ground organization," *Psychological Bulletin*, vol. 138, no. 6, pp. 1172–1217, 2012.

[40] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. V. Fua, and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 2274–2282, 2012. [Online]. Available: https://doi.org/10.1109/TPAMI.2012.120

[41] L. J. Grady, "Random walks for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1768–1783, 2006.

[42] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," in *Neural Information Processing Systems*, 2003.

[43] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Neural Information Processing Systems*, 2013.

[44] X. Li, T. Wei, Y. P. Chen, Y.-W. Tai, and C.-K. 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)*, 2020, pp. 2869–2878. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Li_FSS-1000_A_1000-Class_Dataset_for_Few-Shot_Segmentation_CVPR_2020_paper.html

[45] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. V. Gool, "The 2017 DAVIS Challenge on Video Object Segmentation," *ArXiv*, vol. abs/1704.00675, 2017. [Online]. Available: https://arxiv.org/abs/1704.00675

[46] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. 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)*. IEEE, 2018, pp. 168–172. [Online]. Available: https://doi.org/10.1109/ISBI.2018.8363547

[47] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, *Kvasir-SEG: A Segmented Polyp Dataset*. Springer International Publishing, 2020, pp. 451–462. [Online]. Available: https://doi.org/10.1007/978-3-030-37734-2_37

[48] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, pp. 283–293, 2014.

[49] J. 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. 23 605–23 615, 2024. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Herzog_Adapt_Before_Comparison_A_New_Perspective_on_Cross-Domain_Few-Shot_Segmentation_CVPR_2024_paper.html

[50] Y. Ye, K. Xu, Y. Huang, R. Yi, and Z. Cai, "DiffusionEdge: Diffusion probabilistic model for crisp edge detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 6675–6683.

[51] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 21 875–21 911. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/hash/26cfdcd8fe6fd75cc53e92963a656c58-Abstract-Conference.html

[52] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[53] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, pp. 167–181, 2004.