Supplementary Information for 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

Supplementary Methods 1

Independent Boundary-First implementations

The source-family experiment tests whether the Boundary-First ordering remains effective when query boundaries are supplied in substantially different ways. Each implementation first constructs closed hypotheses from an appearance, geometry, superpixel or gradient source and only then evaluates support-conditioned identity. SAM2 is not loaded in any of these experiments. The six rows are locked, independently evaluated implementations of a common Boundary-First solver rather than one-variable ablations of the principal UBET system. They share a frozen DINOv2 representation, a boundary-aware feature graph, closed-hypothesis generation, candidate scoring and geometric refinement. The boundary provider and its associated identity initialisation are fixed for an entire evaluation; neither is selected using the query annotation. No backbone or boundary detector is fitted on an evaluation set.

Provider maps

Region construction uses a common 448×448 working grid, with DINOv2 features aligned to a 48×48 grid. DiffusionEdge uses the frozen `BSDS_sample.yaml` configuration and `bsds.pt` checkpoint. The depth provider uses the frozen Depth Anything V2 Small checkpoint (`Depth-Anything-V2-Small-hf` from `depth-anything`); predicted depth is bicubically resized, min–max normalised and converted into a discontinuity map by the magnitude of horizontal and vertical 3×3 Sobel derivatives. The fused implementation uses $E_{\text{diff}} + 0.4E_{\text{depth}}$, followed by min–max normalisation.

SLIC is applied with 120 target segments, compactness 10, Gaussian smoothing $\sigma = 1$ and zero-based labels. Felzenszwalb partitioning uses scale 120, $\sigma = 0.8$ and minimum region size 40. For either partition, horizontal and vertical label transitions are marked and dilated once with a 3×3 kernel to produce a boundary map. Canny uses grayscale thresholds 40 and 140. No learned boundary or depth model is loaded in the SLIC, Felzenszwalb or Canny implementations.

Support-conditioned identity evidence

Frozen DINOv2 features are extracted independently from the support and query images. Foreground patches within the support mask and background patches outside it define the identity evidence. In the DiffusionEdge-only and depth-only implementations, transport-based initialisation relates these support patch banks to query patches. The fused DiffusionEdge–depth, SLIC, Canny and Felzenszwalb implementations instead use the difference between mean foreground and background prototypes. A sparse feature graph links each query patch to spatial neighbours and its nearest semantic neighbours; affinities that cross a strong provider boundary are attenuated. The resulting support-conditioned field supplies semantic evidence for region formation and candidate ranking.

Supplementary Table 1: Locked configurations of the independent Boundary-First source family. Each source constructs closed candidates before support-conditioned identity ranking. No row uses SAM2.

Implementation	Native representation	Identity initialisation	Learned source
Diffusion edge + depth edge	Fused edge map	Mean prototypes	Yes
Diffusion edge	Edge probability	Patch transport	Yes
Depth edge	Depth discontinuity	Patch transport	Yes
SLIC superpixels	Label transitions	Mean prototypes	No
Canny edges	Binary edge map	Mean prototypes	No
Felzenszwalb regions	Label transitions	Mean prototypes	No

Closed-hypothesis population

Each provider generates a population of alternatives rather than one thresholded mask. The provider map is thresholded at 0.10, 0.15, 0.20, 0.25 and 0.30. Elliptical closing kernels of size 5, 9 and 13 convert open edge fragments into region barriers; connected components of the complement are filled to form closed hypotheses. Regions containing at most 30 pixels or at least 95% of the image are removed, as are regions with insufficient support-conditioned evidence. These hypotheses are combined with regions obtained from the semantic and depth fields, deduplicated geometrically and, when necessary, reduced to the 50 most plausible candidates before detailed scoring. SLIC and Felzenszwalb enter this same conversion through their label transition maps, whereas Canny enters through its binary gradient map. Query annotations never enter construction, filtering or ranking.

Candidate arbitration

Every retained region is evaluated using support-to-query feature similarity, foreground–background contrast, semantic-field agreement, depth consistency when available and alignment with the provider boundary. The highest-ranked closed region is geometrically refined and returned as the prediction. These operations use only the support episode, the query image and the provider-generated candidates. Ground-truth query masks are accessed only after prediction for evaluation, and episodes are processed independently.

No provider or threshold is selected for an individual query. These implementations

therefore support portability of the Boundary-First decision order, not numerical equivalence among providers or a strict provider-only ablation. The main manuscript reports their complete quantitative results.

Supplementary Methods 2

Exact UBET region-evidence definitions

Let Q_v be a normalised query token, S^+ the support-foreground token bank, $\bar{S}^+$ and $\bar{S}^-$ the foreground and exterior-ring prototypes, and $d(v) = \langle Q_v, \bar{S}^+ \rangle - \langle Q_v, \bar{S}^- \rangle$. Define $h(v)$ as the weighted sum of the within-map ranks of $d(v)$ and $\max_{s \in S^+} \langle Q_v, s \rangle$, with weights 0.55 and 0.45. The binary seed map is $z(v) = \mathbf{1}[h(v) \geq Q_{0.85}(h)]$. For candidate r, the primitive scores are

$$
\begin{aligned}
p_{\text{proto}} &= \langle \overline{Q_r}, \bar{S}^+ \rangle, & p_{\text{set}} &= |r|^{-1} \sum_{v \in r} \max_{s \in S^+} \langle Q_v, s \rangle, \\
p_{\text{rev}} &= |S^+|^{-1} \sum_{s \in S^+} \max_{v \in r} \langle s, Q_v \rangle, & p_{\text{con}} &= |r|^{-1} \sum_{v \in r} d(v), \\
p_{\text{pur}} &= |r|^{-1} \sum_{v \in r} \mathbf{1}[d(v) > 0], & p_{\text{bnd}} &= \overline{d(r)} - \overline{d(E(r))}, \\
p_{\text{match}} &= |C(r)|^{-1} \sum_{v \in C(r)} \max_{s \in S^+} \langle Q_v, s \rangle.
\end{aligned}
$$

Here $C(r)$ and $E(r)$ are the inner contour and exterior ring defined in the main Methods. The bidirectional identity term is

$$
\begin{aligned}
p_{\text{bi}} = 0.12\mathcal{R}(p_{\text{proto}}) + 0.22\mathcal{R}(p_{\text{set}}) + 0.27\mathcal{R}(p_{\text{rev}}) + 0.13\mathcal{R}(p_{\text{con}}) \\
+ 0.09\mathcal{R}(p_{\text{pur}}) + 0.05\mathcal{R}(p_{\text{bnd}}) + 0.12\mathcal{R}(p_{\text{match}}).
\end{aligned}
$$

Masked square crops of the support and candidate provide patch-token and class-token cosines p_{patch} and p_{cls}, giving $p_{\text{crop}} = 0.65 p_{\text{bi}} + 0.15\mathcal{R}(p_{\text{patch}}) + 0.20\mathcal{R}(p_{\text{cls}})$. The released

Supplementary Table 2: Fixed decision paths on identical frozen evidence. Values are foreground mIoU percentages. The best and second-best values in each row are shown in bold and underlined, respectively.

Dataset	n	Boundary only	Semantic only	Fixed intersection	Full UBET
PerSeg	176	<u>93.71</u>	92.78	92.19	**94.68**
DAVIS	1,969	79.27	<u>80.58</u>	78.41	**80.83**
ISIC	650	61.71	61.33	<u>63.43</u>	**65.52**
Kvasir-SEG	200	63.12	<u>66.51</u>	63.28	**70.14**

coverage-and-closure score is

$$a_{\mathrm{cov}} = 0.30p_{\mathrm{crop}} + 0.25\mathcal{R}(p_{\mathrm{seedcov}}) + 0.15\mathcal{R}(p_{\mathrm{mass}}) + 0.15\mathcal{R}(p_{\mathrm{seedprec}}) + 0.10\mathcal{R}(p_{\mathrm{match}}) + 0.05\mathcal{R}(p_{\mathrm{pur}}),$$

where $p_{\mathrm{seedcov}} = \sum_{v\in r} z(v)/\sum_{v\in V} z(v)$, $p_{\mathrm{seedprec}} = |r|^{-1}\sum_{v\in r} z(v)$ and $p_{\mathrm{mass}} = \sum_{v\in r} h(v)$. Denominators are lower-bounded by 10^{-6} in implementation. All ranks are computed among candidates from the same query and no score is carried between episodes.

Supplementary Results

Complete decision-path controls

The main-text intervention reverses the order of boundary and identity decisions. Supplementary Table 2 provides additional fixed-path controls on the same frozen evidence. No control regenerates proposals or accesses a query annotation. The conditional full rule is strongest on all four domains, while the relative ranking of the fixed alternatives changes by domain.

Frozen identity-feature controls

DINOv2 ViT-B/14 and ViT-L/14 were evaluated separately while image preprocessing, SAM2 settings, score definitions and the conflict rule were held fixed. The locked B+L representation is used for every headline result; no backbone is selected per dataset. Because DINOv2 also supplies prompt locations, changing the feature scale may change

Supplementary Table 3: Frozen DINOv2 feature-scale controls on the complete declared evaluation partitions. ViT-B/14, ViT-L/14 and the locked B+L configuration are paired controls. The best and second-best values in each row are shown in bold and underlined, respectively.

Dataset	n	ViT-B/14	ViT-L/14	B+L
PerSeg	176	91.56	**94.70**	<u>94.68</u>
DAVIS	1,969	80.63	<u>80.76</u>	**80.83**
ISIC	650	**66.13**	65.25	<u>65.52</u>
Kvasir-SEG	200	68.37	<u>69.63</u>	**70.14**

Supplementary Table 4: Sensitivity to the shared conflict threshold. Values are foreground mIoU percentages. Range is the maximum minus minimum value across the four thresholds.

Dataset	$\tau = 0.15$	$\tau = 0.25$	$\tau = 0.35$	$\tau = 0.45$	Range
PerSeg	95.1	94.7	94.2	95.0	0.9
DAVIS	79.9	80.8	80.4	80.5	0.9
ISIC	66.7	65.5	66.2	64.2	2.5
Kvasir-SEG	69.3	70.1	69.9	69.1	1.0
FSS-1000	87.5	87.2	87.3	86.4	1.1

the prompted candidate pool.

Conflict-threshold sensitivity

The conflict threshold is shared across datasets. Across the four tested values, the mIoU range is 0.9 points on PerSeg, 0.9 on DAVIS, 2.5 on ISIC, 1.0 on Kvasir-SEG and 1.1 on FSS-1000. Sensitivity is modest on four datasets and greater on ISIC. The predeclared value $\tau = 0.25$ is nevertheless retained throughout the main evaluation to avoid dataset-specific selection.

Boundary–content dissociation

Region-specific perturbations probe whether UBET responds only to interior appearance or only to contour information. Perturbations are applied to the query image, while the annotated support image and support mask remain unchanged. The query annotation is used only to construct this retrospective diagnostic and to compute IoU; it is never supplied to the deployable predictor. On the 256×256 working grid, a Euclidean distance

transform orders foreground pixels by distance from the annotated contour. The closest 25% form the boundary region, and the deepest 25% form an equal-area interior region. A Gaussian-blur delta is restricted to the selected region, and its global image-space root-mean-square energy is matched between the boundary and interior arms. Each perturbed query is then processed independently by the complete frozen UBET inference path; clean candidates or predictions are not reused. In the benchmark-level controls, perturbing either region reduces accuracy, and boundary-region perturbation has the larger reduction on all four datasets (Supplementary Table 5). This result supports functional separation of the two evidence paths, not independence of their effects.

Supplementary Table 5: Boundary–content dissociation across four domains. Boundary and Interior denote mIoU after the corresponding region-specific perturbation. Drops are measured relative to Clean; positive Δ indicates the larger loss after boundary perturbation.

Dataset	Clean	Boundary	Interior	Boundary drop	Interior drop	Δ
PerSeg	94.70	90.20	91.94	4.50	2.76	+1.74
DAVIS	80.80	76.30	77.10	4.50	3.70	+0.80
ISIC	65.50	62.10	64.50	3.40	1.00	+2.40
FSS-1000	87.20	84.70	85.30	2.50	1.90	+0.60

The FSS-1000 population stress audit provides a complementary test over 240 queries sampled from the 2,400-image test set, with affected object fraction and perturbation energy matched. Mild corruption has little effect on either path; under severe blur, interior corruption becomes more damaging (Supplementary Table 6). Thus the model is not a contour-only detector: boundaries define candidate extent, whereas interior appearance remains necessary for identity arbitration.

Supplementary Table 6: Matched FSS-1000 population stress test over 240 queries sampled from the 2,400-image test set. Entries are mIoU-point changes from the clean diagnostic mean of 88.55; positive values indicate degradation and a small negative value denotes a small increase.

Gaussian σ	Boundary-region drop	Interior-region drop	Boundary minus interior
3	0.44	-0.18	+0.61
5	1.67	1.66	+0.01
7	1.94	4.27	-2.33
9	2.50	7.43	-4.93

Accuracy, footprint and inference cost

Supplementary Table 7 compares accuracy, parameter footprint and inference time across the evaluated methods. Accuracy is the descriptive mean foreground mIoU across the seven benchmarks. Parameters include all instantiated pretrained weights, and runtime is end-to-end serial query latency after initialisation on one RTX 3090.

Supplementary Table 7: Accuracy, footprint and inference cost. 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	46.13	641.09	0.928
PerSAM-F	54.37	641.09	7.376
FSSDINO	57.78	<u>85.70</u>	**0.138**
PR-MaGIC	69.75	945.46	11.682
INSID3	66.49	303.15	0.7331
Matcher	64.47	945.46	6.244
GF-SAM	68.33	945.46	0.736
ABCDFSS	52.97	**25.56**	2.821
FSS-SAM3	71.06	840.51	<u>0.698</u>
UBET (SAM2)	<u>75.87</u>	615.40	2.481
UBET (SAM3)	**75.95**	1,231.46	1.8146

Supplementary Table 8 reports the runtime attribution of UBET. The timing values constitute a practical common-hardware profile rather than an architecture-intrinsic speed ranking because released repositories expose different profiling interfaces.

Supplementary Table 8: Runtime attribution of UBET. Values are the mean stage times and shares from the 100-episode RTX 3090 instrumented trace used for the principal UBET efficiency profile in the main text.

Stage	Mean time (s)	Runtime share (%)
SAM2 proposal generation	1.012	40.79
Dense support localisation	0.217	8.74
Closed-group construction	0.330	13.29
Identity transport and scoring	0.922	37.17
Conflict certificate	< 0.001	0.01

Supplementary reproducibility statement

All experiments use one annotated support example, frozen foundation encoders and the declared protocol of the corresponding main-text analysis. Episodes are evaluated independently, and no query annotation enters inference. Empty predictions are retained. Query annotations are used only for retrospective metrics and explicitly labelled oracle or perturbation audits. The inference and evaluation code is publicly available through the GitHub repository cited in the main manuscript.