BEYOND SELECTION: INTEGRATING COMPLEMENTARY INFORMATION FOR TRAINING-FREE VISUAL IN-CONTEXT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent visual foundation models provide strong visual representations and segmentation capabilities, enabling training-free visual prompt-based in-context segmentation (VP-ICS) for novel segmentation tasks. However, existing VP-ICS methods predominantly rely on winner-take-all selection strategies, which retain only a limited subset of retrieved evidence and can discard complementary information distributed across multiple support shots. We propose Integra-ICS, a training-free VP-ICS framework that integrates support evidence to exploit such distributed information. For each query token, Integra-ICS retrieves the global top-k support tokens from a unified support token pool containing both foreground and background. By applying masked softmax aggregation to query-support token similarities, Integra-ICS integrates complementary evidence across support shots. Candidate foreground regions are subsequently localized and filtered by a minimum-evidence criterion to obtain reliable predictions. Across eight diverse benchmarks, Integra-ICS outperforms the strongest baseline by 3.60 and 3.56 mIoU points on average in the 5-shot and 10-shot settings, respectively, using a single hyperparameter configuration without dataset-specific tuning. The improvement comes from integrating complementary cross-shot evidence, while the minimum-evidence criterion retains reliable predictions supported by foreground tokens from the support set.

1 INTRODUCTION

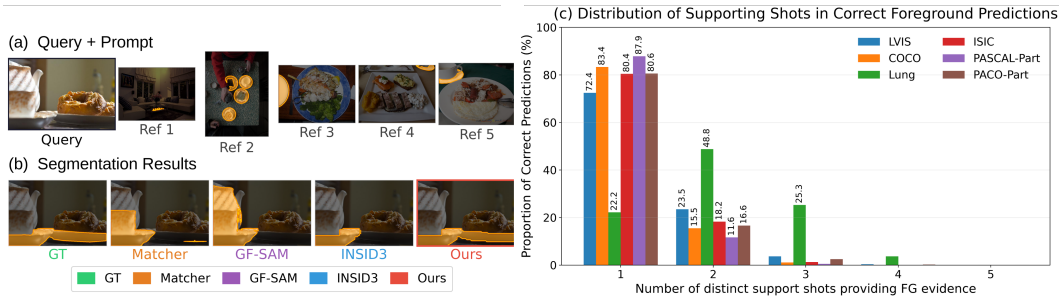

Figure 1: Qualitative illustration and motivation of the proposed Integra-ICS. (a) An example of few-shot segmentation with a query image and five reference support shots. (b) Qualitative comparison against baseline methods, showing that our method produces a segmentation mask closest to the ground truth (GT). (c) Analysis of the number of distinct support shots contributing to correctly predicted foreground tokens during 5-shot inference across datasets. The analysis shows that correct foreground predictions can benefit from complementary evidence distributed across multiple support shots.

Recent foundation models (FMs), such as SAM (Kirillov et al., 2023) and DINOv3 (Siméoni et al., 2025), have significantly advanced the field of image segmentation by providing rich and generalizable representations learned from large-scale data. Because these models provide strong visual

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

representations and, in some cases, segmentation capabilities even with limited data, they have become well suited for adaptation to novel segmentation tasks. However, since fine-tuning these large models to every new task is not a scalable approach, effectively leveraging their pretrained knowledge is becoming increasingly important. Accordingly, training-free **I**n-**C**ontext **S**egmentation (ICS) methods relying solely on visual prompts (VP) have recently been proposed (Liu et al., 2023; Zhang et al., 2024; Cuttano et al., 2026).

These training-free VP-ICS methods typically follow a three-step paradigm. First, a frozen vision encoder extracts dense features from both the reference and target images. Next, foreground features from the reference are used to score the target features based on feature-level similarity, and spurious query correspondences are pruned via a verification step (e.g., backward matching (Cuttano et al., 2026)). Finally, the verified query correspondences are converted into a final mask, either by using them as prompts for segmentation models such as SAM, or by clustering target features and scoring the resulting clusters using query and support-level information (Cuttano et al., 2026).

Despite their promising results, existing VP-ICS frameworks face an information bottleneck in exploiting multiple support shots due to their reliance on winner-take-all (WTA) selection paradigms. While these methods aim to identify reliable foreground evidence, their rigid selection strategies can suppress complementary cues across shots. Specifically:

- Matcher (Liu et al., 2023) establishes one-to-one bipartite matching and retains only a sparse subset of top-ranked points, thereby ignoring multiple supporting cues for a given query region.
- GF-SAM (Zhang et al., 2024) scores each query token using its single most similar foreground support token within each shot. Although the resulting shot-wise maps are subsequently aggregated, all other within-shot correspondences are discarded before aggregation.
- INSID3 (Cuttano et al., 2026) reduces the backward correspondence from each shot to an unmasked 1-NN and converts foreground membership into a binary vote. Query tokens falling below the vote threshold are rejected even when their retrieved similarities provide additional evidence.

As illustrated in Figure 1(c), our analysis reveals the presence of non-negligible complementary information beyond the top-1 retrieved token. Consequently, these rigid WTA selection schemes can prematurely discard such distributed evidence, limiting their ability to fully exploit multi-shot contexts.

To overcome this limitation, we propose Integra-ICS (**Integr**ating support cues for **I**n-**C**ontext **S**egmentation). Rather than restricting correspondences to WTA-selected tokens within individual shots, Integra-ICS pools all support tokens—including both foreground and background regions—into a unified support pool. For each query token, it retrieves the top-k most similar support tokens globally, allowing complementary evidence from multiple shots to be jointly considered. We then apply a softmax to the cosine similarities between the query and the retrieved support tokens, allowing both foreground and background tokens to contribute. This naturally suppresses false positives when foreground support tokens exhibit weak similarity to the query, as background tokens receive relatively higher weights. The softmax weights of the retrieved foreground tokens are then aggregated to obtain a candidate score for each query location. The resulting score map is thresholded to obtain a binary candidate mask, which guides the extraction of query cluster prototypes. These clusters are scored and thresholded to obtain a binary prediction mask. Finally, our minimum-evidence mask is applied via an element-wise AND to obtain the final segmentation. By integrating complementary evidence across multiple support shots and filtering unsupported predictions, Integra-ICS addresses the multi-shot information bottleneck of WTA-based paradigms while improving prediction reliability. Across eight diverse benchmarks spanning general objects (COCO-20^i, LVIS-92^i), fine-grained parts (PASCAL-Part, PACO-Part), remote sensing and underwater domains (iSAID, SUIM), and biomedical imaging (Lung Segmentation, ISIC), our framework achieves average improvements of $+3.60$ and $+3.56$ mIoU points over the strongest baseline in the 5-shot and 10-shot settings, respectively, using a single universal hyperparameter configuration without per-dataset tuning. The primary contributions of this work are summarized as follows:

- **Multi-shot Information Bottleneck**: We identify a limitation of existing training-free VP-ICS methods, which predominantly rely on selection-based mechanisms and may discard complementary foreground evidence across support shots.
- **Integrated Support Evidence and Minimum-Evidence Filtering**: We propose **Integra-ICS**, a training-free framework that integrates complementary evidence from multi-shot support tokens through unified global top-k retrieval and similarity-based soft aggregation, while introducing a minimum-evidence mask to filter unsupported predictions.
- **Cross-Domain Generalizability**: We validate the broad applicability of Integra-ICS across eight diverse benchmarks spanning everyday scenes, fine-grained parts, remote sensing, underwater environments, and biomedical scans. Using a single universal hyperparameter configuration without per-dataset tuning, our framework demonstrates consistent improvements over existing training-free methods.

2 Related Work

Visual In-context Segmentation. Matcher (Liu et al., 2023) establishes an in-context learning framework that leverages pre-trained visual foundation models (VFMs) for general perception tasks. Specifically, it performs patch-level feature matching via VFMs (e.g., DINOv2 (Oquab et al., 2023)) to identify bidirectional correspondences, which serve as point prompts for a promptable mask decoder like SAM (Kirillov et al., 2023). Building upon this two-stage paradigm, GF-SAM (Zhang et al., 2024) enhances prompt quality by constructing positive-negative similarity prior maps across reference and target images. It samples high-confidence visual prompts to produce candidate mask proposals via SAM, subsequently employing graph-based connected components and point consistency verification to filter and aggregate the optimal target masks. Departing from this heavy reliance on auxiliary prompt-guided decoders, INSID3 (Cuttano et al., 2026) eliminates SAM entirely from the pipeline. Instead, it directly exploits the rich, discriminative representations of DINOv3 to achieve accurate segmentation, thereby avoiding prompt-sampling overhead. Meanwhile, extending beyond purely visual cues, CG-ICS (Chen et al., 2026) incorporates textual concept reasoning via multimodal large language models (MLLMs) to disambiguate target categories and guide mask generation.

Unlike existing visual-only methods that rely on rigid winner-take-all (WTA) selection, Integra-ICS integrates complementary evidence across multiple support shots. In contrast to multimodal approaches, it remains strictly within the visual-only regime, alleviating the multi-shot information bottleneck without relying on external linguistic priors or incurring the overhead of heavy MLLM inference.

3 Method

3.1 Feature Extraction

Figure 2 represents our proposed framework. Leveraging the semantically informative representations of DINOv3 (Siméoni et al., 2025), Integra-ICS use DINOv3 vision encoder as its feature extractor. Let N_Q and N_S denote the number of spatial tokens per query and support image, respectively, let N_{shot} denote the number of support shots, and let D be the feature channel dimension. The encoder yields query feature tokens $\mathbf{F}_Q = \{\mathbf{f}_Q^q \in \mathbb{R}^D\}_{q=1}^{N_Q}$ and support feature tokens $\mathbf{F}_S = \{\mathbf{f}_S^{s,j} \in \mathbb{R}^D \mid 1 \le s \le N_{\text{shot}}, 1 \le j \le N_S\}$, accompanied by the binary support masks $\mathbf{M}_S = \{M_S^{s,j} \in \{0,1\} \mid 1 \le s \le N_{\text{shot}}, 1 \le j \le N_S\}$.

To establish a unified semantic reference across multiple shots, we first construct a foreground prototype for each support shot via masked average pooling and ℓ_2 normalization. We then average the shot-level prototypes and apply ℓ_2 normalization again to obtain the global support prototype:

$$\bar{\mathbf{p}}_S^s = \mathrm{Norm}\left(\frac{\sum_{j=1}^{N_S} M_S^{s,j}\mathbf{f}_S^{s,j}}{\sum_{j=1}^{N_S} M_S^{s,j} + \epsilon}\right), \quad \mathbf{p}_S = \mathrm{Norm}\left(\frac{1}{N_{\text{shot}}}\sum_{s=1}^{N_{\text{shot}}} \bar{\mathbf{p}}_S^s\right), \tag{1}$$

where $\mathrm{Norm}(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|_2$ and ϵ prevents division by zero. The extracted representations $\mathbf{F}_Q$, $\mathbf{F}_S$, and masks $\mathbf{M}_S$ are subsequently forwarded to extract the minimum evidence and candidate masks.

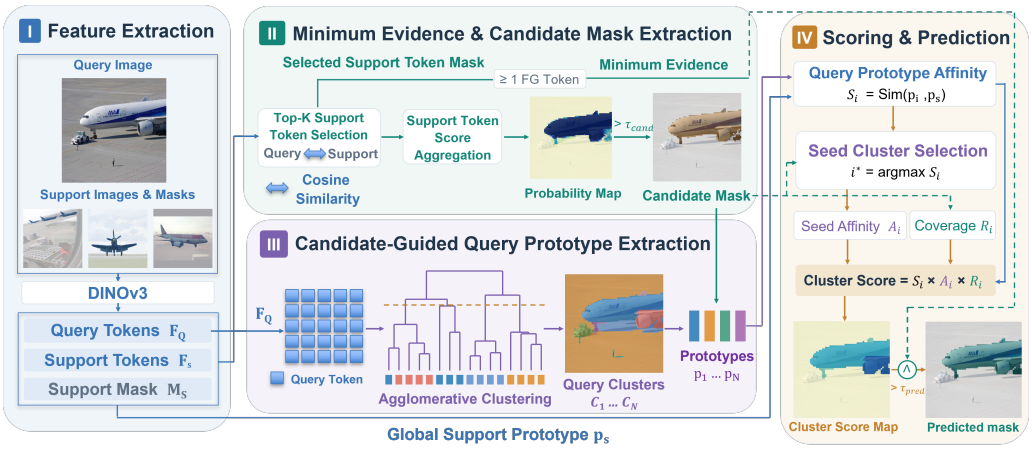

Figure 2: **Overview of Integra-ICS**. Integra-ICS consists of four stages: (1) DINOv3 feature extraction, (2) candidate mask and minimum evidence extraction from query-support correspondences, (3) prototype extraction utilizing the candidate mask, and (4) final prediction through semantic, contextual and coverage-based cluster scoring with minimum-evidence filtering.

3.2 Minimum Evidence & Candidate Mask Extraction

Let $T_S = N_{\text{shot}} N_S$ denote the total number of support tokens. Pairwise correlation between flattened support features $\tilde{\mathbf{F}}_S \in \mathbb{R}^{T_S \times D}$ and query features $\mathbf{F}_Q \in \mathbb{R}^{N_Q \times D}$ is evaluated via cosine similarity, constructing the token affinity matrix:

$$\mathbf{S}^{\text{tok}} = \text{Sim}(\tilde{\mathbf{F}}_S, \mathbf{F}_Q) \in \mathbb{R}^{T_S \times N_Q}, \tag{2}$$

where $S^{\text{tok}}_{m,q}$ represents the semantic correlation between the m-th support token and the q-th query token ($1 \le m \le T_S, 1 \le q \le N_Q$). For each query token q, we retrieve its top-K highest similarity scores $\mathbf{s}_q = [s_{1,q}, \ldots, s_{K,q}]^\top \in \mathbb{R}^K$ along with their corresponding support ground-truth mask labels $\tilde{\mathbf{m}}_q = [\tilde{M}_S^{1,q}, \ldots, \tilde{M}_S^{K,q}]^\top \in \{0,1\}^K$, where $\tilde{M}_S^{k,q}$ denotes the foreground/background label of the k-th retrieved support token for query token q. We define the *minimum evidence* criterion to identify query tokens that receive at least one foreground-supported correspondence from the support set. Formally, we count the number of foreground tokens among the top-K retrieved support tokens:

$$V_q = \sum_{k=1}^{K} \tilde{M}_S^{k,q}. \tag{3}$$

A query token satisfies the minimum-evidence condition if at least one of its retrieved support tokens belongs to the foreground:

$$M^q_{\text{min}} = \mathbb{I}\left(V_q > 0\right) = \mathbb{I}\left(\sum_{k=1}^{K} \tilde{M}_S^{k,q} \ge 1\right), \tag{4}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Thus, $M^q_{\text{min}} = 0$ indicates that all of its retrieved support tokens are background. While semantic clustering groups visually coherent regions, it is inherently susceptible to semantic drift (i.e., erroneously expanding into visually similar background). To prevent unconstrained false-positive leakage, we introduce the minimum evidence as a safeguard. Meanwhile, we compute a foreground evidence score $E_q \in [0,1]$ via masked softmax aggregation:

$$w_{k,q} = \frac{\exp(s_{k,q})}{\sum_{k'=1}^{K} \exp(s_{k',q})}, \quad E_q = \sum_{k=1}^{K} w_{k,q} \tilde{M}_S^{k,q}, \tag{5}$$

The binary candidate mask $M_{\text{cand}} \in \{0,1\}^{N_Q}$ is then obtained by thresholding E_q with τ_{cand}:

$$M^q_{\text{cand}} = \mathbb{I}\left(E_q \ge \tau_{\text{cand}}\right). \tag{6}$$

M_{cand} is used to guide prototype extraction in Section 3.3 and to compute the coverage ratio in Section 3.4.

3.3 CANDIDATE-GUIDED QUERY PROTOTYPE EXTRACTION

Merely comparing query and support features to obtain a similarity map often confines the activation to localized regions with dominant similarity. To address this limitation, Matcher and GF-SAM leverage SAM to delineate semantically coherent foreground regions. Also, INSID3 employs agglomerative clustering to identify foreground clusters and derive representative prototype vectors. To delineate semantically coherent regions without prompt-based decoders, we group query tokens into clusters $\{C_i\}_{i=1}^{N}$ via agglomerative clustering. Since agglomerative clustering groups semantically similar query tokens, the resulting clusters may include query tokens that are not directly supported by the retrieved support tokens. Therefore, the minimum-evidence mask ensures that only query tokens with at least one foreground support token among their top-K retrieved tokens are retained in the final prediction. To refine prototype extraction using the candidate foreground regions, we use the candidate mask to select query tokens for prototype aggregation:

$$\mathbf{p}_i = \frac{\sum_{q \in C_i} M^q_{\text{cand}} \mathbf{f}^q_Q}{\sum_{q \in C_i} M^q_{\text{cand}} + \epsilon}. \tag{7}$$

3.4 SCORING & PREDICTION

Equipped with the global support prototype $\mathbf{p}_S$ and candidate-guided query prototypes $\{\mathbf{p}_i\}_{i=1}^{N}$, we evaluate cross-image semantic affinity for each cluster:

$$S_i = \text{Sim}(\mathbf{p}_i, \mathbf{p}_S), \quad 1 \le i \le N. \tag{8}$$

The cluster exhibiting the highest similarity is designated as the seed cluster, indexed by $i^* = \arg\max_i S_i$. This seed guides the computation of intra-image contextual affinity $A_i = \text{Sim}(\mathbf{p}_i, \mathbf{p}_{i^*})$. Concurrently, we measure the spatial reliability of each cluster by calculating its coverage ratio R_i with respect to the candidate mask:

$$R_i = \frac{\sum_{q \in C_i} M^q_{\text{cand}}}{|C_i|}. \tag{9}$$

Combining these criteria, each cluster is assigned a composite score:

$$\text{Score}_i = S_i \times A_i \times R_i. \tag{10}$$

Finally, each query token is identified as foreground if its associated cluster score exceeds the prediction threshold τ_{pred} and it satisfies the minimum evidence criterion, yielding the final prediction mask $M_{\text{pred}} \in \{0,1\}^{N_Q}$:

$$M^q_{\text{pred}} = \mathbb{I}\left(\text{Score}_{i(q)} > \tau_{\text{pred}}\right) \cdot M^q_{\text{min}}, \tag{11}$$

where $i(q)$ denotes the cluster index to which query token q belongs.

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

We evaluate our method across eight few-shot segmentation datasets: LVIS-92^i (Liu et al., 2023), COCO-20^i (Nguyen & Todorovic, 2019), ISIC (Codella et al., 2019), iSAID (Yao et al., 2021), SUIM (Islam et al., 2020), X-Ray (Candemir et al., 2013), PASCAL-PART (Liu et al., 2023), and PACO-PART (Liu et al., 2023). These datasets span diverse domains, including everyday scenes, satellite imagery, underwater environments, medical imaging, and part segmentation. For all experiments, we set k=5 for top-k retrieval, τ_{cand}=0.4, and τ_{pred}=0.2. Unless explicitly stated otherwise, these default settings are used across all evaluations. In addition, agglomerative clustering uses cosine distance with average linkage, with a distance threshold fixed at 0.25 for all experiments. The hyperparameter settings used for the baseline methods are described in the Appendix.

4.2 MAIN RESULTS

We conducted comparative experiments against four baselines—Matcher, GF-SAM, INSID3, and RSRM (Sun et al., 2026)—across all eight benchmarks. As shown in Table 1, Integra-ICS achieves

Table 1: **Main benchmark results on 5-shot and 10-shot visual in-context segmentation.** All methods are evaluated training-free using frozen vision backbones. † denotes our re-evaluation using the official codebase without CRF post-processing. Best results are highlighted in **bold**. All evaluations use a single universal configuration without per-dataset tuning.

Method	Backbone	Semantic						Part		Avg.
		LVIS-92^i	COCO-20^i	ISIC	iSAID	SUIM	X-Ray	PASCAL	PACO	
5-Shot Evaluation										
Matcher (Liu et al., 2023)	DINOv2	40.00	60.70	35.00	34.30	50.60	71.20	45.80	33.60	46.40
GF-SAM (Zhang et al., 2024)	DINOv2	44.20	**66.80**	55.20	52.40	58.10	52.90	51.20	41.90	52.80
INSID3† (Cuttano et al., 2026)	DINOv3	45.87	64.68	63.80	58.12	62.60	78.10	56.01	46.55	59.47
RSRM (Sun et al., 2026)	DINOv3	44.79	63.90	62.55	56.92	63.28	**87.01**	58.58	49.89	60.87
Integra-ICS (Ours)	DINOv3	**54.86**	65.90	**66.13**	**63.55**	**64.35**	86.72	**62.00**	**52.22**	**64.47**
10-Shot Evaluation										
Matcher (Liu et al., 2023)	DINOv2	41.66	62.06	37.45	39.46	48.18	71.16	38.60	35.77	46.79
GF-SAM (Zhang et al., 2024)	DINOv2	46.17	67.23	59.32	55.64	66.05	48.91	49.93	43.67	54.62
INSID3† (Cuttano et al., 2026)	DINOv3	47.07	64.83	65.84	59.05	64.54	77.92	57.58	50.2	60.86
RSRM (Sun et al., 2026)	DINOv3	44.79	65.00	65.31	58.99	64.77	**90.09**	60.67	53.07	62.84
Integra-ICS (Ours)	DINOv3	**56.76**	**67.85**	**67.35**	**67.68**	**66.48**	86.49	**63.42**	**55.19**	**66.40**

the highest mIoU on most benchmarks under both 5-shot and 10-shot settings, with X-Ray remaining the exception and COCO-20^i favoring GF-SAM in the 5-shot setting. In particular, Integra-ICS outperforms RSRM, the strongest baseline on average, by 3.60 and 3.56 mIoU points in the 5-shot and 10-shot settings, respectively. The consistent gains across both settings and diverse datasets indicate that Integra-ICS effectively exploits complementary information from multiple support examples without training or task-specific adaptation.

Figure 3 presents qualitative comparisons with the baseline methods. RSRM tends to produce overly broad predictions and fails to distinguish fine-grained semantic differences between foreground and surrounding regions, as observed in the first and third rows. Matcher and GF-SAM also exhibit incomplete foreground coverage in the fourth row, where several target instances are missed. Although INSID3 identifies a larger portion of the foreground, it still fails to recover some target regions. In contrast, our method provides more complete foreground predictions across the examples by exploiting complementary evidence from multiple support instances. Overall, these qualitative results are consistent with the quantitative improvements of Integra-ICS, particularly in capturing fine-grained semantic differences.

4.3 Complementary Evidence Rescue

To quantify how effectively complementary evidence rescues false negatives at the token level, we design the cross-shot complementary rescue rate (CSCR). Let FN_1 denote the query-FG tokens whose top-1 support neighbor is a background token, and let $\mathrm{Rec} \subseteq \mathrm{FN}_1$ denote those recovered by the top-k softmax aggregation $m(q) > \tau_{cand}$ (k=5, τ_{cand}=0.4). CSCR measures the fraction of recovered tokens whose remaining neighbors contain at least two FG tokens originating from at least two distinct support shots.

$$\mathrm{CSCR} = \frac{|q \in \mathrm{Rec} : |F(q)| \geq 2,\ |S(q)| \geq 2|}{|\mathrm{Rec}|}. \tag{12}$$

$F(q)$ denotes the set of foreground support tokens among the top-k retrieved neighbors of query token q, excluding the top-1 match, while $S(q)$ denotes the set of distinct support shots containing the tokens in $F(q)$.

Table 2 reports the CSCR and the fraction of top-1 false-negative tokens recovered by complementary evidence. The Top-1 FN Ratio denotes the fraction of false-negative tokens among all query-FG tokens, while the Rescue Rate is defined as $|\mathrm{Rec}|/|\mathrm{FN}_1|$. The Complementary Recovery Rate, computed as the product of the Rescue Rate and CSCR, represents the fraction of top-1 false negatives ultimately rescued through cross-shot complementary evidence. Table 2 shows that top-1 matching produces false negatives across all eight benchmarks, with an average FN ratio of 20.73%. Importantly, top-k aggregation rescues a non-negligible fraction of these false negatives on every



Figure 3: **Qualitative comparison with baseline methods**. Results on LVIS-92^i with 5-shot inference. Blue regions in the query column denote the ground-truth masks, while green regions in the remaining columns denote the predicted masks of each method.

Table 2: Token-level analysis of false-negative rescue under 5-shot inference. We report the ratio of query-FG tokens whose top-1 support neighbor is a background token, the fraction of these false negatives rescued by top-k softmax aggregation ($k = 5$), the cross-shot complementary rescue rate (CSCR), and the complementary recovery rate.

Dataset	Top-1 FN Ratio (%)	Rescue Rate (%)	CSCR (%)	Complementary Recovery Rate (%)
PACO-Part	29.45	9.19	21.97	2.02
PASCAL-Part	20.50	11.90	15.16	1.80
LVIS-92^i	19.62	13.28	29.66	3.94
iSAID	21.01	13.92	22.28	3.10
ISIC	26.63	8.18	20.62	1.69
SUIM	24.09	10.90	22.54	2.46
X-Ray	6.02	25.92	74.39	19.28
COCO-20^i	18.52	16.38	21.39	3.51
Average	20.73	13.71	28.50	4.73

benchmark, with an average rescue rate of 13.71%. Among the rescued tokens, 28.50% exhibit cross-shot complementary evidence, yielding an average complementary recovery rate of 4.73%. These results indicate that complementary foreground evidence is consistently present across different benchmarks and can recover query foreground tokens that are missed by top-1 matching. The substantial variation across datasets further suggests that the availability of such complementary evidence depends on the underlying visual characteristics of each benchmark. In particular, X-Ray exhibits a notably high CSCR of 74.39%, indicating that many false negatives can be recovered with complementary information from distinct shots. Overall, these findings support our hypothesis that complementary evidence distributed across multiple shots can provide information beyond the single best match.

4.4 Effect of Top-K and the Number of Support Shots

As shown in Figure 4, increasing top-k consistently improves mIoU across different support-shot settings on three of the four benchmarks, with X-Ray being the exception. In contrast, X-Ray achieves its best performance at k=3 and exhibits a gradual performance drop as k increases further, as a larger k makes the minimal-evidence (ME) criterion increasingly permissive. This suggests that

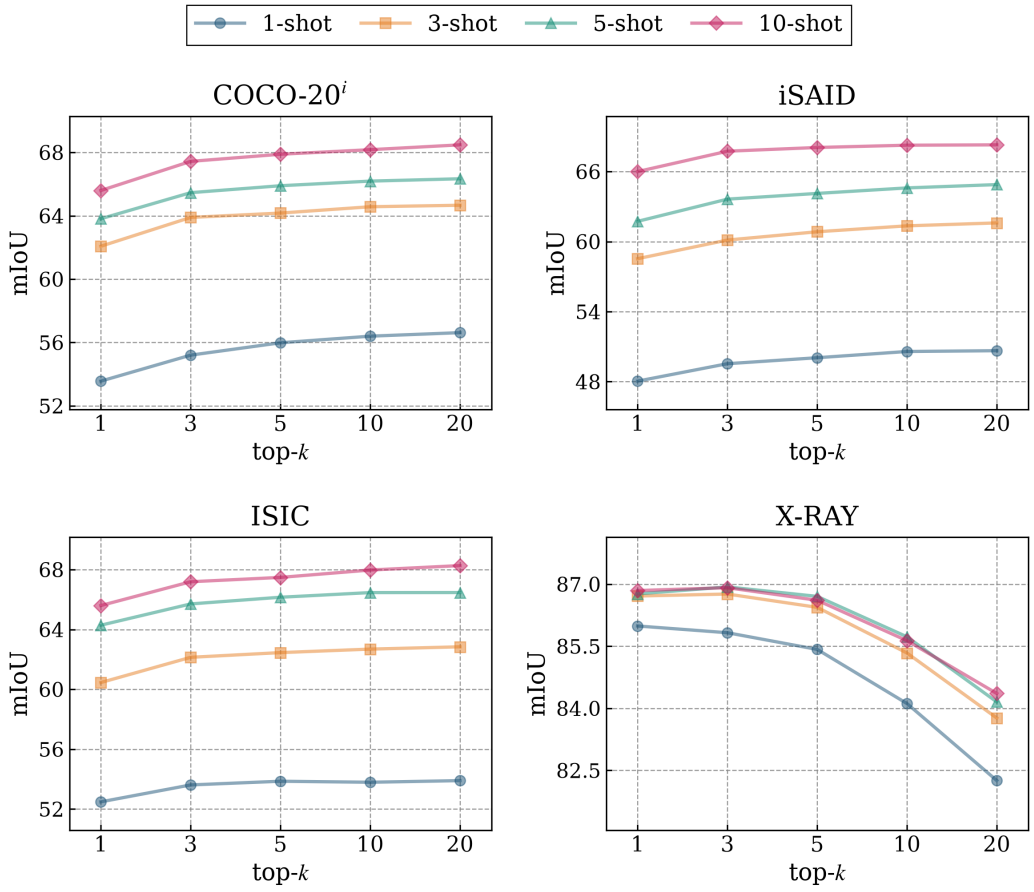

Figure 4: Performance of Integra-ICS with varying top-k and numbers of support shots across four representative benchmarks from diverse domains.

the optimal k depends on the benchmark. A detailed analysis of the performance degradation on X-Ray as top-k increases is provided in the Appendix. Meanwhile, the performance trends with respect to k remain consistent across different numbers of support shots, indicating that the effectiveness of complementary evidence is not limited to a specific support-shot setting. Overall, these results show that complementary evidence can be effectively exploited across different numbers of support shots, while the optimal k varies with benchmark characteristics.

4.5 Ablation Study and Threshold Analysis

Ablation Study. We conduct an ablation study to examine the effects of minimal evidence (ME) and candidate mask-guided prototype extraction (CMPE). As shown in Table 3, removing either component leads to a decrease in the average performance, although their contributions differ substantially in magnitude. Removing minimal evidence (**ME**) causes a notable performance drop of 1.82 mIoU on average, whereas removing **CMPE** leads to drop of 0.16 mIoU. In detail, CMPE acts to reduce false negatives while increasing false positives; since the reduction in false negatives marginally exceeds the increase in false positives, it leads to an overall net gain (detailed in Appendix).

The impact of ME is particularly pronounced on X-Ray, where removing it decreases performance by 8.14 mIoU. This indicates that ME effectively suppresses unreliable predictions arising from semantic clustering. In contrast, on COCO-20^i and iSAID, removing ME slightly improves performance, as ME occasionally suppresses valid foreground regions expanded through clustering when their query tokens lack direct foreground support. This is an inherent side effect of the safeguard mechanism. Qualitative examples illustrating these effects are provided in the Appendix.

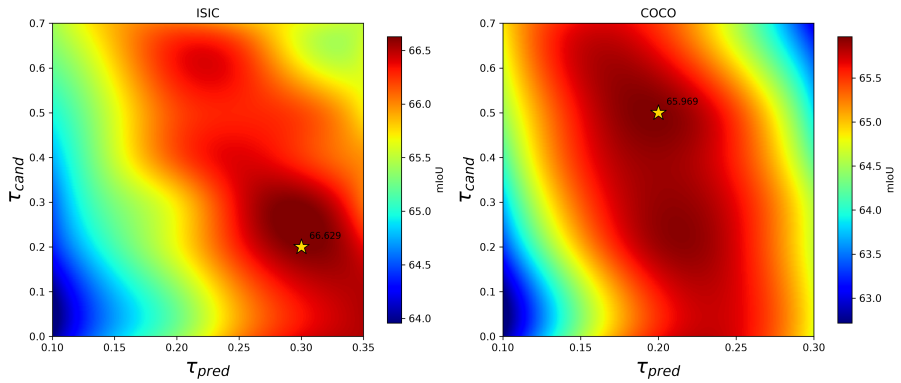

Figure 5: Sensitivity analysis of τ_{cand} and τ_{pred} based on mIoU. ISIC and COCO-20^i are used as representative datasets under the 5-shot inference setting. Stars indicate the best-performing configuration for each dataset

Table 3: Ablation study under 5-shot inference, with results reported in mIoU (%). Each variant removes the indicated component from the full framework. **ME** and **CMPE** denote minimal evidence and candidate mask-guided prototype extraction, respectively. Note that ME serves as a **safeguard**, whereas CMPE is designed to enhance recall.

	COCO-20^i	PASCAL-Part	iSAID	ISIC	X-RAY	Avg.
Full	65.90	62.00	63.55	66.13	86.72	68.86
− ME	66.26(+0.36)	60.95(-1.05)	64.08(+0.53)	65.32(-0.81)	78.58(-8.14)	67.04(-1.82)
− CMPE	65.92(+0.02)	61.91(-0.09)	63.25(-0.30)	66.01(-0.12)	86.39(-0.33)	68.70(-0.16)
− ME, CMPE	66.58(+0.68)	61.45(-0.55)	63.85(+0.30)	65.58(-0.55)	79.21(-7.51)	67.33(-1.53)

As shown in Table 3, removing candidate mask-guided prototype extraction (CMPE) leads to performance drops across most benchmarks (up to -0.33 mIoU on X-Ray), with an average drop of 0.16 mIoU. In Eq.7, CMPE uses the candidate mask M_{cand} to restrict query prototype aggregation to reliable foreground regions. In addition, removing both ME and CMPE yields 67.33 mIoU on average, which is higher than removing ME alone (67.04 mIoU) but remains below the full framework (68.86 mIoU). This indicates that ME and CMPE interact non-additively, where ME provides the primary safeguard and CMPE offers an additional refinement.

Threshold Analysis. We conduct experiments to examine the sensitivity of performance to the choice of τ_{cand} and τ_{pred}. Figure 5 visualizes the mIoU landscape over the two thresholds on ISIC and COCO-20^i. Although the best-performing configurations vary across the two datasets, both exhibit relatively broad high-performance regions. In particular, τ_{pred}=0.2 consistently falls within a high-performance region across both datasets.

5 Conclusion

In this paper, we identify a limitation of existing training-free visual in-context segmentation methods: they make limited use of information distributed across support shots when establishing query-support correspondences. To address this limitation, we introduce **Integra-ICS**, which integrates complementary evidence from multiple support shots to recover foreground tokens potentially missed by sparse correspondence selection. To improve prediction reliability, we further employ minimal evidence filtering to suppress unsupported predictions. Extensive experiments on eight benchmarks demonstrate that Integra-ICS consistently outperforms the strongest baseline by an average of 3.60 and 3.56 mIoU points under the 5-shot and 10-shot settings, respectively. Our analyses reveal that integrating complementary evidence and minimal evidence filtering contribute to overall performance.

AI USE STATEMENT

We used generative AI tools to assist with manuscript refinement and research-related tasks. This included identifying potential omissions or logical inconsistencies in the manuscript, discussing possible additional experiments and analyses, and brainstorming ideas to improve the completeness and rigor of the work. We also used AI tools to assist in developing experimental code for implementing selected analyses and experiments.

For manuscript preparation, we used AI tools to improve the structure, organization, flow, and readability of the initial and revised drafts. We reviewed all AI-generated suggestions and code, and independently checked and validated the resulting experimental implementations and results. We take full responsibility for the final content, methodology, experiments, claims, code, and conclusions of this work.

REFERENCES

Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2):577–590, 2013.

Zhigang Chen, Xiawu Zheng, and Rongrong Ji. Toward robust in-context segmentation via concept guidance. *arXiv preprint arXiv:2606.28149*, 2026.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

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*, pp. 21638–21648, 2026.

Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 1769–1776. IEEE, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*, 2023.

Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *2019 IEEE/CVF international conference on computer vision (ICCV)*, pp. 622–631. IEEE, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Sujun Sun, Mingwu Ren, and Haofeng Zhang. Training-free cross-domain few-shot segmentation via robust semantic representation and matching. *arXiv preprint arXiv:2606.24297*, 2026.

Xiwen Yao, Qinglong Cao, Xiaoxu Feng, Gong Cheng, and Junwei Han. Scale-aware detailed matching for few-shot aerial image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021.

Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi H Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. *Advances in Neural Information Processing Systems*, 37:33232–33261, 2024.

A APPENDIX

Inference Speed Analysis.Table 4 compares the inference runtime across training-free visual in-context segmentation frameworks under the 5-shot setting at an image resolution of 518×518. All comparisons are conducted on a single H100 GPU with an Intel Xeon Processor Model 143. Methods relying on prompt-guided mask decoders (e.g., Matcher and GF-SAM) suffer from substantial computational overhead due to candidate sampling and SAM decoding iterations, yielding less than 1.0 FPS. In contrast, by eliminating auxiliary prompt decoders, Integra-ICS operates at 1.945 FPS, running significantly faster than RSRM (1.504 FPS) and closely matching the speed of INSID3 (2.028 FPS). While our framework extends 1-NN retrieval to top-k evidence integration across a unified support pool, this additional step introduces negligible latency relative to feature extraction. Consequently, Integra-ICS runs at 1.945 FPS, close to INSID3 at 2.028 FPS and faster than the SAM-based Matcher and GF-SAM under the same 5-shot, 518×518 setup.

Table 4: **Inference speed comparison** under the 5-shot setting with an input resolution of 518×518. All methods are measured on the same hardware without task-specific tuning.

Method	**Backbone**	**Inference Speed (FPS)** ↑
Matcher (Liu et al., 2023)	DINOv2	0.236
GF-SAM (Zhang et al., 2024)	DINOv2	0.855
RSRM (Sun et al., 2026)	DINOv3	1.504
INSID3 (Cuttano et al., 2026)	DINOv3	**2.028**
Integra-ICS (Ours)	DINOv3	<u>1.945</u>

One-Shot Inference Evaluation. Our framework is primarily designed to maximize performance in the multi-shot scenario by combining complementary information across distinct support shots. Nevertheless, since our framework naturally supports one-shot inference, we additionally evaluate its performance in the one-shot setting. As shown in Table 5, our approach achieves the highest performance on average among the compared methods. For a fair comparison with the results in Table 1, INSID3 is evaluated without applying Conditional Random Field (CRF) post-processing.

Table 5: Performance comparison under 1-shot setting. Results are reported in mIoU (%).

Method	COCO	LVIS	Lung	ISIC	SUIM	iSAID	PACO-Part	Pascal-Part	Avg.
Ours	55.98	**44.45**	85.42	53.86	**53.74**	**49.30**	**42.44**	**52.57**	**54.72**
RSRM	53.23	36.10	**88.77**	**56.76**	53.75	48.40	39.27	49.19	53.18
INSID3	56.09	41.12	77.40	55.40	52.90	48.80	38.15	48.17	52.25
GF-SAM	**58.70**	35.20	51.00	48.70	53.10	47.10	36.30	44.50	46.83
Matcher	52.70	33.00	70.80	38.60	44.10	33.30	34.70	42.90	43.76

Candidate Mask-Guided Prototype Extraction. Figure 6(a) shows, across three datasets, the difference between the increase in true positives and the increase in false positives when applying CMPE compared with the setting without CMPE. Except for the 10-shot setting on ISIC, the increase in false positives is consistently smaller than the increase in true positives. Furthermore, Figure 6(b) shows that CMPE consistently reduces false negatives across all three datasets and shot settings. These results support our observation that CMPE is designed to prioritize improving recall.

The Effect of Minimum Evidence on Foreground Expansion. We further analyze the marginal performance decrease caused by applying ME on datasets such as COCO-20^i. As shown in Figure 7, qualitative comparisons reveal that ME can remove foreground regions that are effectively expanded through semantic clustering but lack direct support from the retrieved support tokens. This suggests that, while ME helps suppress unsupported predictions, its support requirement can also discard valid foreground regions recovered through the clustering process, resulting in a marginal performance degradation on some datasets.

Table 6: 5-shot inference results on the X-Ray dataset with minimal evidence constructed using only the top-1 retrieved token.

Top-k	3	5	10
mIoU (%)	86.82	87.08	87.15

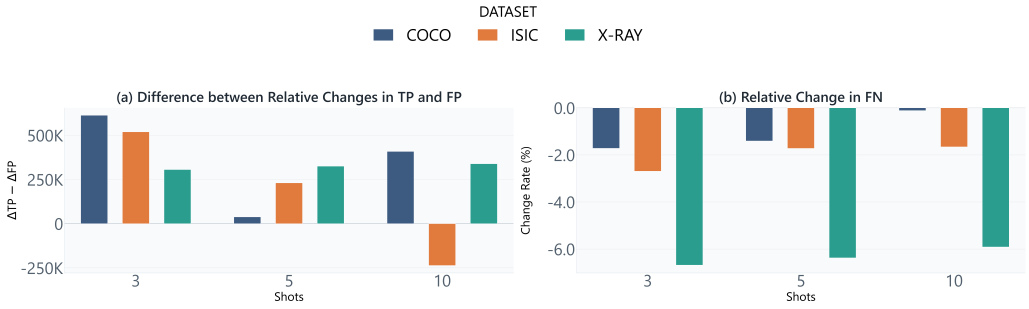

Figure 6: **Candidate Mask-guided Prototype Extraction Contribution Analysis.** (a) shows the difference between the changes in true positives and false positives, computed as $\Delta TP - \Delta FP$. Here, ΔTP and ΔFP denote the changes in true positives and false positives, respectively, after applying CMPE. (b) shows the change in false negatives after applying CMPE across three datasets and three few-shot settings (3-, 5-, and 10-shot).

Minimum Evidence and Its Permissive Tendency as Top-k Increases. Table 6 reports the 5-shot inference results on the X-Ray dataset, where $K{=}1$ is fixed when constructing minimal evidence (ME) following Eq. (4), while $k{=}3, 5, 10$ is used for the softmax aggregation of complementary evidence. As shown in the table, performance consistently improves as k increases, similar to the trend observed on the other datasets. This indicates that, as discussed in the main text, ME can become increasingly permissive depending on the dataset.

Hyperparameters of baselines. For all baseline experiments, we use the hyperparameters reported by the baseline methods for obtaining their main results. For INSID3, however, we do not apply CRF post-processing, and set the distance threshold for agglomerative clustering to 0.25, the same value used by our method, to ensure a fair comparison.

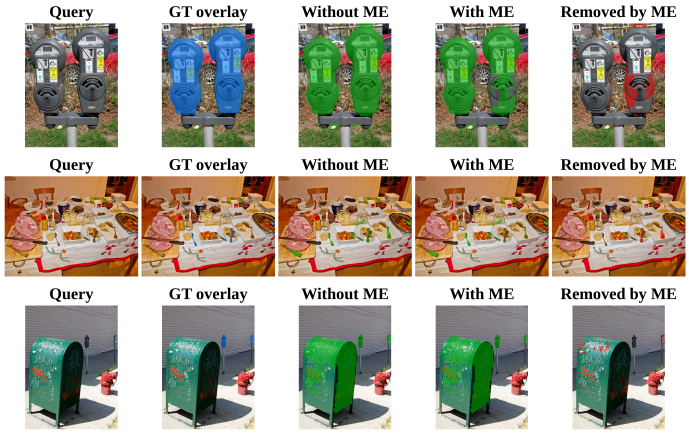

Figure 7: **The Effect of Minimum Evidence on Foreground Expansion.** Qualitative comparisons between two cases: with and without minimum evidence. Green regions denote predicted foreground regions, while red regions in the last column highlight the differences between the two cases. Results are obtained under the 5-shot inference setting on the COCO-20^i dataset.