RELIABLE EVIDENCE LEARNING FOR SMALL TARGETS IN UNIFIED CONTEXT-DEPENDENT CONCEPT SEGMENTATION

Anonymous authors
Paper under double-blind review

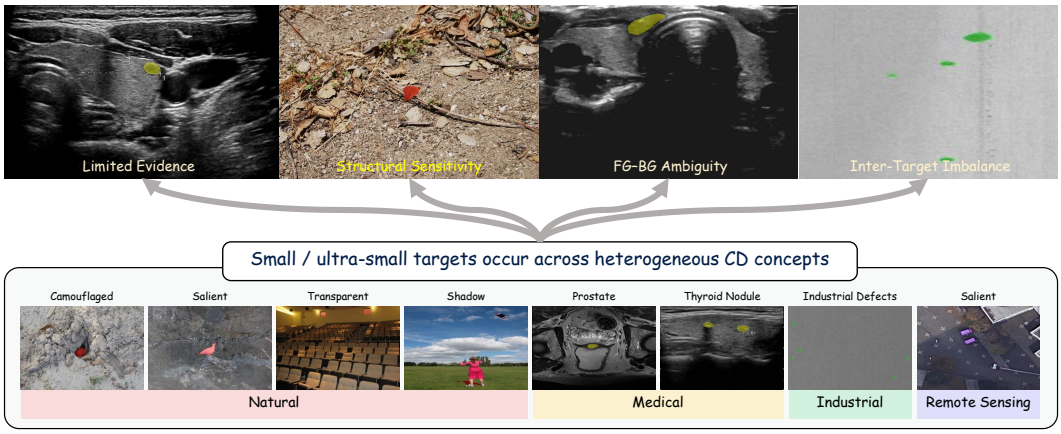

Figure 1: Small and ultra-small targets across context-dependent (CD) concepts. Top: shared challenges of limited spatial evidence, structural sensitivity, foreground–background ambiguity, and inter-target imbalance. Bottom: representative concepts spanning natural, medical, industrial, and remote-sensing scenarios.

ABSTRACT

Unified context-dependent concept segmentation aims to segment heterogeneous visual concepts with a shared model conditioned on support examples. Small targets challenge this paradigm because their sparse visual evidence is easily overwhelmed by background structures, while locally strong responses may still be inconsistent with the specified concept. We present RELIC, a framework for reliable evidence learning that distributes concept conditioning across representation adaptation, evidence verification, and final prediction. Specifically, Size- and Concept-aware Small-target Evidence Verification (SC-SEV) matches query features against foreground and background support descriptor groups and combines relative foreground–background evidence with an absolute foreground criterion. The verification criteria adapt to target occupancy and support–query agreement, allowing weak but concept-consistent responses to receive stronger emphasis. To refine the remaining errors, an Evidence-Guided Small-Target Mixture-of-Experts (EGST-MoE) combines query-level expert weighting with spatial gating for weak-response recovery, boundary refinement, fragment completion, and false-positive suppression. Without bells and whistles, RELIC achieves consistently strong performance across eight benchmarks spanning natural, industrial, remote-sensing, and medical scenarios, with particularly pronounced advantages in segmenting small and ultra-small targets.

1 INTRODUCTION

Unified segmentation seeks to accommodate heterogeneous visual tasks within a shared model, enabling knowledge sharing while reducing the need for task-specific architectures. This goal is par-

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.

ticularly challenging for context-dependent (CD) concepts, whose foreground definitions depend on relationships between a target and its surrounding context. Examples include salient and camouflaged objects, shadows, transparent surfaces, industrial defects, and medical lesions. Recent unified approaches use visual demonstrations, support representations, or task prototypes to specify these heterogeneous concepts within a common segmentation framework (Zhao et al., 2024b; Chang et al., 2025; Xu et al., 2026b; Zhao et al., 2026). Although these approaches provide effective mechanisms for concept specification and shared representation learning, segmenting small targets remains difficult because the visual evidence needed to identify the requested foreground is spatially limited and easily overwhelmed by competing structures.

As shown in Fig. 1, small and ultra-small targets occur across diverse CD concepts in natural, medical, industrial, and remote-sensing scenarios. Despite their different appearances and semantic meanings, these targets share several intrinsic difficulties. Limited evidence provides few discriminative cues, making target responses vulnerable to feature downsampling and background interference. Structural sensitivity means that small localization errors can substantially alter a target's shape or remove it from the prediction. Foreground–background ambiguity further complicates discrimination when targets resemble surrounding structures, while inter-target imbalance allows larger foreground regions to dominate learning. These challenges have motivated high-resolution modeling, cross-scale interaction, local refinement, and size-aware optimization (Lou et al., 2022; Kong et al., 2024; Li et al., 2024). Under unified modeling, however, preserving a small response is only useful if that response is consistent with the concept specified by the support examples.

Concept information can condition shared representations at the input (Wang et al., 2023a;b) or guide foreground–background discrimination at the tail (Zhao et al., 2024b). For small targets, input-level conditioning does not explicitly verify concept consistency through subsequent feature transformations, whereas tail-only conditioning may act after useful local evidence has been attenuated. The upper-left region of Fig. 2 illustrates this limitation of conditioning at a restricted stage. Although recent methods condition both encoding and decoding (Xu et al., 2026b), the reliability of local responses under sparse visual evidence remains a separate concern. A support set specifies the intended concept, but even a strong query activation may reflect a competing background structure. This creates an evidence gap between concept specification and reliable small-target prediction, motivating explicit verification of local responses against both foreground and background support evidence.

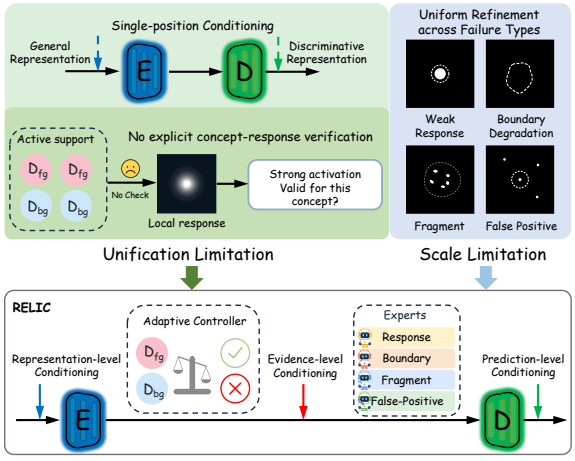

Figure 2: Coupled limitations of unified small-target CD concept segmentation and the design of RELIC.

Small-target evidence exhibits heterogeneous failures, including weak responses, degraded boundaries, fragmented regions, and false positives (Fig. 2, upper right). These failures require different corrections: response enhancement may amplify false positives, while boundary recovery and fragment completion require different cues. Uniform refinement may therefore be insufficient when failures coexist. Together, the unification and scale limitations motivate determining *which responses are consistent with the active concept* and *how and where unreliable evidence should be refined*.

In this work, we present RELIC, a framework for reliable evidence learning that addresses concept consistency and small-target evidence degradation through distributed conditioning, adaptive verification, and state-guided refinement (Fig. 2, lower region). **First,** we distribute support-based concept conditioning across representation adaptation, evidence verification, and final prediction. Support-derived concept representations guide hierarchical feature adaptation, foreground and background descriptors provide verification references, and a dynamic concept filter conditions the final prediction. **Second,** we introduce Size- and Concept-aware Small-target Evidence Verification (SC-SEV), which matches query locations against multiple exemplars within foreground and background support descriptor groups. Dual verification combines relative foreground–background evidence with an absolute foreground criterion, adapting to target occupancy and support–query

agreement to emphasize weak but concept-consistent responses while accounting for competing background evidence. **Third,** we introduce an Evidence-Guided Small-Target Mixture-of-Experts (EGST-MoE) with four experts for weak-response recovery, boundary refinement, fragment completion, and false-positive suppression. Query-state-guided weighting controls each expert's contribution, while expert-specific spatial gates determine where corrections are applied. SC-SEV verifies concept-relative evidence, while EGST-MoE uses the evidence states to guide targeted refinement.

Our main contributions are summarized as follows:

- We propose RELIC, a unified framework that addresses concept consistency and small-target evidence degradation through concept conditioning across representation adaptation, evidence verification, and final prediction.
- We introduce SC-SEV to verify local query responses against foreground and background support descriptor groups using dual criteria adapted to target occupancy and support–query agreement.
- We develop EGST-MoE to address heterogeneous small-target failures by combining query-state-guided expert weighting with expert-specific spatial gating for targeted refinement.
- Experiments across eight benchmarks demonstrate that RELIC achieves consistently strong performance across CD concepts, with pronounced advantages on small and ultra-small targets.

2 Related Work

2.1 Context-Dependent Concept Segmentation

Context-dependent (CD) concept segmentation encompasses tasks whose foreground interpretation relies on surrounding context, including salient and camouflaged objects (Wang et al., 2017; Fan et al., 2020a), shadows (Vicente et al., 2016), industrial defects (Cui et al., 2023), remote-sensing targets (Tu et al., 2022), and medical structures and lesions (Litjens et al., 2014; Gong et al., 2023). Existing methods address these tasks through contextual modeling, multi-scale representation, boundary refinement, and feature interaction (Pang et al., 2022; 2024; He et al., 2023; Lu et al., 2025; Fan et al., 2020b;c). Although effective within their respective domains, these approaches are typically trained for individual tasks with fixed foreground definitions, limiting opportunities for knowledge sharing across heterogeneous CD concepts. Unified modeling requires a shared representation to support different foreground interpretations according to the requested concept. Small targets make this requirement challenging because limited spatial evidence weakens concept discrimination, as reflected in the degradation of generalist segmentors on extremely small CD targets (Zhao et al., 2024a). RELIC focuses on this intersection of concept variation and spatial sparsity, assessing whether local query evidence is consistent with the concept specified by support examples.

2.2 Unified Vision Models

Unified vision models seek to share representations across heterogeneous tasks while retaining the ability to follow different task specifications. Painter represents visual tasks through image-form inputs and outputs (Wang et al., 2023a), while SegGPT unifies segmentation tasks through in-context coloring (Wang et al., 2023b). UniverSeg uses support image-mask pairs to specify medical segmentation tasks through interactions between query and support features (Butoi et al., 2023). For CD concepts, Spider introduces group prompts and a concept filter to obtain concept-specific predictions from shared representations (Zhao et al., 2024b). Subsequent studies explore self-referring indicators (Chang et al., 2025), task-conditioned adaptation and prototype-guided decoding (Xu et al., 2026b), and rule-induced concept grounding (Zhao et al., 2026). These approaches advance task specification and shared representation learning through different forms of conditioning. RELIC complements this line of research by explicitly assessing whether sparse query responses are supported by the active concept and connecting this assessment to targeted refinement.

2.3 Small Object Perception

Small-object perception spans detection, tracking, re-identification, and segmentation, where limited spatial support weakens semantic discrimination and localization precision. Detection methods exploit high-resolution features through sparse computation (Yang et al., 2022) or introduce point prompts to strengthen localization and semantic cues for tiny instances (Zhu et al., 2026). Track-

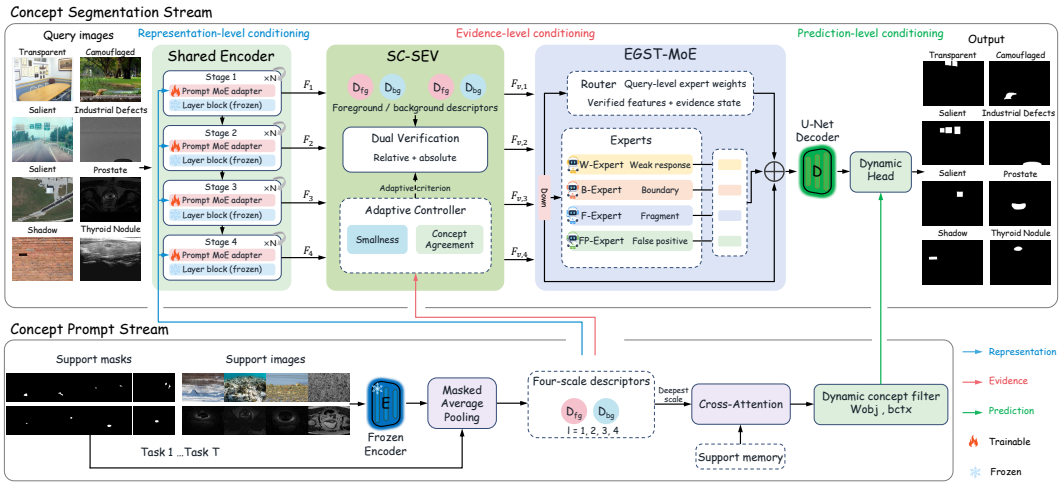

Figure 3: Overview of RELIC. The concept prompt stream constructs foreground/background support descriptor groups to condition the segmentation stream at representation, evidence, and prediction levels. Prompt MoE adapters adapt query features, SC-SEV verifies local evidence, and EGST-MoE performs state-guided refinement before decoding and dynamic concept prediction.

ing and cross-resolution re-identification improve discriminative representations through multi-level knowledge distillation (Zhu et al., 2024) and resolution-invariant learning (Li et al., 2019), respectively. In segmentation, contextual aggregation, cross-scale interaction, and feature preservation improve small-target representations (Lou et al., 2022; Kong et al., 2024; Dai et al., 2026), while complementary approaches incorporate imaging-physics priors (Yang et al., 2026) and size-invariant optimization (Li et al., 2024). These works strengthen small-object representations across tasks. Unified CD concept segmentation further requires sparse responses to remain consistent with the concept specified by support examples, motivating our concept-relative evidence learning.

3 METHOD

Framework Overview. Given a query image I_q and a support set $\mathcal{S} = \{(I_s^n, M_s^n)\}_{n=1}^{N_s}$, RELIC predicts the foreground specified by the support image–mask pairs. As shown in Fig. 3, following the common practice of recent unified frameworks that leverage strong pretrained backbones to adapt to diverse visual concepts (Xu et al., 2026b; Zhao et al., 2024b; Chang et al., 2025; Zhao et al., 2026), the concept segmentation and concept prompt streams share a frozen SAM 2 (Ravi et al., 2025) backbone. The query branch incorporates Prompt MoE adapters and extracts four-scale representations $\{F_l\}_{l=1}^4$, with $F_l \in \mathbb{R}^{C_l \times H_l \times W_l}$. The support branch constructs foreground and background descriptor groups $\{\mathcal{D}_{\mathrm{fg},l}, \mathcal{D}_{\mathrm{bg},l}\}_{l=1}^4$ through masked average pooling. These descriptors provide concept conditioning at three levels. For representation adaptation, the normalized foreground–background contrast $c_l = \mathrm{Norm}(\mathrm{Mean}(\mathcal{D}_{\mathrm{fg},l}) - \mathrm{Mean}(\mathcal{D}_{\mathrm{bg},l}))$ conditions the adapters at stage l. For evidence verification, SC-SEV evaluates local query responses, and EGST-MoE produces refined features $\{F_l'\}_{l=1}^4$. For final prediction, a shared U-Net decoder fuses these features into F_{dec}. The deepest support descriptors interact with the support memory through cross-attention to generate a dynamic concept filter $\langle W_{\mathrm{obj}}, b_{\mathrm{ctx}} \rangle$, which produces $\mathrm{Logits} = \mathrm{DynamicConv}(F_{\mathrm{dec}}, W_{\mathrm{obj}}) + b_{\mathrm{ctx}}$, where W_{obj} and b_{ctx} denote the dynamic kernel and contextual bias.

Size- and Concept-aware Small-target Evidence Verification (SC-SEV). SC-SEV preserves support descriptor groups to assess query evidence against multiple foreground and background exemplars. At query location p, the matching response is computed as:

$$A_{r,l}(p) = \mathrm{TopKMean}_{d \in \mathcal{D}_{r,l}} \cos\left(\phi_{q,l}(F_l(p)), \phi_{r,l}(d)\right), \qquad r \in \{\mathrm{fg}, \mathrm{bg}\}, \tag{1}$$

where $\phi_{q,l}$ and $\phi_{r,l}$ are learnable projections, and TopKMean averages the K largest similarities. The resulting maps describe foreground support and competing background evidence.

To condition verification on target occupancy, we derive a support cue ρ_s from the foreground ratios of the support masks and a soft query estimate $\rho_q = \mathrm{StopGrad}(\mathrm{Mean}(V_{\mathrm{pre},1}))$, where $V_{\mathrm{pre},1} = \sigma((A_{\mathrm{fg},1} - \bar{\alpha} A_{\mathrm{bg},1})/\bar{\tau})$ uses the finest-scale responses and fixed reference coefficients.

The occupancy cues produce a learned smallness state:

$$s = \sigma\left(\mathrm{MLP}\left([-\log(\rho_s + \epsilon), -\log(\rho_q + \epsilon)]\right)\right), \tag{2}$$

where σ is the sigmoid function, $\epsilon > 0$ ensures numerical stability, and $\mathrm{StopGrad}$ blocks gradients through the query-occupancy estimate. We further summarize support–query agreement as $r_{c,l} = \sigma(\mathrm{TopKMean}_p(A_{\mathrm{fg},l} - A_{\mathrm{bg},l})/\tau_c)$. The two states adapt verification as follows:

$$\begin{aligned} \alpha_l &= \alpha_{\max}\sigma(a_l - \eta_l s - \xi_l r_{c,l}), \quad & \theta_l &= \theta_l^0 - \delta_l s - \zeta_l r_{c,l}, \\ G_l &= A_{\mathrm{fg},l} - \alpha_l A_{\mathrm{bg},l}, & V_l &= \sigma\left(\frac{G_l}{\tau_{\mathrm{rel},l}}\right)\sigma\left(\frac{A_{\mathrm{fg},l} - \theta_l}{\tau_{\mathrm{abs},l}}\right), \end{aligned} \tag{3}$$

where $\alpha_{\max} > 0$ bounds the background coefficient, θ_l^0 is the base foreground threshold, and $a_l, \eta_l, \xi_l, \delta_l, \zeta_l$ are learnable controller parameters. All temperatures are positive. The two factors in V_l combine relative foreground–background dominance with an absolute foreground-response criterion. The verified evidence modulates the query features as $F_{v,l} = F_l \odot (1 + \gamma_l V_l)$, where γ_l controls the modulation strength and V_l is broadcast across channels.

Evidence-Guided Small-Target Mixture-of-Experts (EGST-MoE). To refine heterogeneous errors after evidence verification, EGST-MoE introduces four experts with complementary refinement biases. The weak-evidence expert E_w uses multi-receptive-field depth-wise convolutions, the boundary expert E_b uses directional and high-frequency filtering, the fragment expert E_f uses dilated context aggregation, and the false-positive expert E_{fp} uses Haar frequency cues for suppression. Query-level weighting and expert-specific spatial gating coordinate these experts across queries and locations. For each scale, spatial pooling gives $u_l = [\mathrm{Mean}(F_{v,l}), \mathrm{Max}(F_{v,l})]$. An evidence-state vector $e_l \in \mathbb{R}^7$ summarizes evidence extent, peak response, local variation, ambiguity, local inconsistency, foreground–background margin, and concept agreement. A learned router jointly encodes $\{u_l, e_l\}_{l=1}^4$ to predict the expert weights:

$$\boldsymbol{\pi} = \mathrm{Softmax}\left(R_\theta(\{u_l, e_l\}_{l=1}^4)\right), \qquad \boldsymbol{\pi} = [\pi_w, \pi_b, \pi_f, \pi_{fp}], \tag{4}$$

where R_θ is the learnable router. All four experts remain active, sharing weights across scales. For reduced inputs $X_l = \mathrm{Down}_l(F_{v,l})$, the three enhancement experts compute gated residuals as:

$$U_{e,l} = \mathrm{Up}_l\left(E_e(X_l) \odot \Gamma_{e,l}\right), \qquad e \in \{w, b, f\}, \tag{5}$$

where $\Gamma_{e,l} \in [0, 1]$ is an expert-specific spatial gate and Up_l restores the residual to the shape of $F_{v,l}$. The false-positive expert suppresses competing responses using a bounded attenuation map:

$$U_{fp,l} = -F_{v,l} \odot B_{fp,l}, \tag{6}$$

where $B_{fp,l}$ combines the expert output with its spatial gate and is broadcast across feature channels. The four residuals refine the verified feature through query-level weighting:

$$F_l' = F_{v,l} + \sum_{e \in \{w,b,f,fp\}} \pi_e U_{e,l}, \tag{7}$$

where F_l' denotes the refined representation. § B defines the evidence state and spatial gates.

Training and Inference. We adopt task-balanced sampling and joint optimization (Zhao et al., 2024b). For each task, we sample N image–mask pairs, using B as queries and the remaining $N_s = N - B$ as support examples. The training objective is defined as:

$$\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_{\mathrm{seg}}^t + \lambda_{\mathrm{rank}}\mathcal{L}_{\mathrm{rank}} + \lambda_{\mathrm{bal}}\left(\mathcal{L}_{\mathrm{enc-bal}} + \mathcal{L}_{\mathrm{ref-bal}}\right), \tag{8}$$

where T is the number of jointly trained tasks. The segmentation loss combines weighted binary cross-entropy and weighted IoU losses (Wei et al., 2020; Fan et al., 2020b) with Dice loss (Milletari et al., 2016). Motivated by size-invariant object balancing (Li et al., 2024), $\mathcal{L}_{\mathrm{rank}}$ applies object-balanced evidence margin supervision to reduce the dominance of large foreground regions, with details provided in § C. The routing-balancing terms discourage expert collapse in the two MoE components. $\mathcal{L}_{\mathrm{enc-bal}}$ follows the standard expert-balancing objective for sparse MoE routing (Shazeer et al., 2017), while $\mathcal{L}_{\mathrm{ref-bal}}$ balances the batch-wise routing importance of the dense EGST-MoE experts, as detailed in § D. The coefficients λ_{rank} and λ_{bal} control the auxiliary objectives. At inference, K-means clustering over encoder features selects 64 representative training image–mask pairs per task. This fixed support set supplies the descriptors and dynamic concept filter for all test queries of that task.

Table 1: Dataset splits and object statistics for eight CD concept segmentation benchmarks. Object size is measured relative to image area. Counts are cumulative: ultra-small ($< 1\%$) objects are included in the small ($< 10\%$) subset, and both subsets are included in all objects.

Dataset	Concept	Scene	Images train + test	Object area ratio (%)	Number of objects Ultra-small $< 1\%$	Small $< 10\%$	All $< 100\%$
COD10K (Fan et al., 2020a)	Camouflaged	Natural	4,040 + 2,026	< 0.001–80.738	16,052	20,169	22,295
DUTS (Wang et al., 2017)	Salient	Natural	10,548 + 5,017	< 0.001–84.590	1,973	6,833	18,893
SBU Shadow (Vicente et al., 2016)	Shadow	Natural	4,085 + 638	< 0.001–87.033	9,037	14,877	17,297
GSD (Lin et al., 2021)	Transparent	Natural	3,285 + 813	< 0.001–99.886	3,016	10,353	13,651
PROMISE12 (Litjens et al., 2014)	Prostate	Medical	778 + 418	0.109–14.418	125	1,175	1,196
TN3K (Gong et al., 2023)	Thyroid nodule	Medical	2,879 + 614	0.020–75.884	268	2,302	3,823
ORSI-4199 (Tu et al., 2022)	Salient	Remote sensing	2,000 + 2,199	< 0.001–80.042	2,017	5,562	7,236
ESDIs-SOD (Cui et al., 2023)	Industrial defects	Industrial	3,600 + 1,200	0.003–85.030	16,778	21,979	23,346
Total	–	–	**31,215 + 12,925**	–	**49,266**	**83,250**	**107,737**

4 Experiments

4.1 Settings

Datasets and Metrics. We evaluate RELIC on eight benchmarks spanning natural, industrial, remote-sensing, and medical scenarios, with training/test splits and object statistics summarized in Tab. 1. For scale-wise evaluation, we characterize each test image by the area of its largest four-connected ground-truth foreground component relative to the image area. Following prior small-target evaluation thresholds (Dai et al., 2026), the ultra-small (U) and small (S) subsets contain images with area ratios below 1% and 10%, respectively, while A denotes the complete test set. These subsets are cumulative, with U included in S and both included in A. We report S_m (Fan et al., 2017), SI-MAE, SI-$F_{\max}$ (Li et al., 2024), and mIoU. The SI metrics follow the object-wise evaluation principle (Li et al., 2024), using foreground connected components as object proxies. Detailed image grouping and metric aggregation procedures are provided in § G.

Implementation Details. We initialize the image encoder for query and support images with the Hiera-L pretrained weights from SAM 2. The pretrained backbone remains frozen, while the adapters and newly introduced modules are optimized. RELIC is jointly trained across eight tasks for 15 epochs using task-balanced sampling on 8 NVIDIA RTX 4090 GPUs. All input images are resized to 512×512. We use Adam (Kingma & Ba, 2015) with an initial learning rate of 1×10^{-4}, multiplied by 0.9 every 10 epochs. At inference, each task uses a fixed support set of 64 training image-mask pairs selected by K-means clustering over encoder features.

4.2 Evaluation

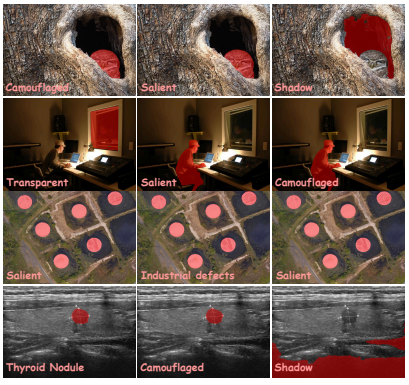

Figure 4: Visualization of multi-concept prediction.

Quantitative Results. Tab. 2 compares RELIC with recent specialized and unified models across eight segmentation benchmarks. For a fair comparison, the input-resolution settings of the unified baselines are detailed in § A. RELIC achieves the best or tied-best results in 67/72 settings (93.1%), covering 24/24 ultra-small and 24/24 small settings (100% each). Compared with Spider, the strongest unified baseline in macro-averaged SI-$F_{\max}$, RELIC achieves relative gains of 2.16% (U), 1.98% (S), and 1.51% (A) across the eight datasets, highlighting its greater advantages at smaller scales. On the ultra-small subset of GSD, RELIC reaches $S_m = 0.948$, improving over the strongest unified baseline by 4.75% and the strongest specialized baseline by 4.52%. These results demonstrate consistently strong performance across heterogeneous CD concepts, particularly when target evidence is spatially limited. Detailed threshold-wise comparisons with unified methods are provided in § H.

Table 2: Comparison with specialized and unified methods across eight benchmarks. U ($< 1\%$), S ($< 10\%$), and A (all) denote cumulative test-image subsets. **Red**, green, and blue indicate the three best distinct scores per row; ties share a rank. Best / Reported counts best or tied-best scores over reported entries. †PROMISE12 has identical S and A subsets.

Evaluation Setting			Specialized Methods								Unified Methods				
			ESCNet	VSCodev2	WEFT	CKLNet	HASNet	SAM2-RS	ACO-MoE	DifferSeg	Spider	SR-ICL	TP-Seg	ConceptSeg	RELIC
Dataset	**Metric**	**Scale**	(Ye et al., 2025)	(Lao et al., 2026b)	(Sun et al., 2026b)	(Shen et al., 2026)	(Ma et al., 2026)	(Sun et al., 2026a)	(Lao et al., 2026a)	(Zhou et al., 2026)	(Zhao et al., 2024b)	(Chang et al., 2025)	(Xu et al., 2026b)	(Zhao et al., 2026)	(Ours)
COD10K	$S_m \uparrow$	U	.731	.736	–	–	–	–	–	–	**.743**	.702	.713	**.754**	**.764**
		S	.851	.850	–	–	–	–	–	–	**.857**	.822	.817	**.862**	**.879**
		A	.869	.870	–	–	–	–	–	–	**.875**	.845	.839	**.887**	**.894**
	SI-MAE ↓	U	.058	.060	–	–	–	–	–	–	**.057**	.062	.065	**.055**	**.054**
		S	.056	.057	–	–	–	–	–	–	.057	.060	**.054**	**.053**	**.050**
		A	**.054**	**.056**	–	–	–	–	–	–	**.056**	.058	.061	**.049**	**.049**
	SI-$F_{\max}$ ↑	U	.712	.739	–	–	–	–	–	–	**.740**	.679	.693	**.758**	**.776**
		S	.806	**.821**	–	–	–	–	–	–	.818	.761	.769	**.832**	**.855**
		A	.834	**.847**	–	–	–	–	–	–	.844	.796	.801	**.853**	**.876**
DUTS	$S_m \uparrow$	U	–	**.840**	.826	–	–	–	–	–	.787	.726	.809	**.829**	**.847**
		S	–	**.915**	**.900**	–	–	–	–	–	.889	.836	.898	**.915**	**.924**
		A	–	**.928**	.916	–	–	–	–	–	.909	.861	.911	**.933**	**.935**
	SI-MAE ↓	U	–	**.018**	**.019**	–	–	–	–	–	.033	.041	**.018**	.029	**.015**
		S	–	**.025**	.028	–	–	–	–	–	.033	.042	**.024**	**.021**	**.021**
		A	–	**.026**	.027	–	–	–	–	–	.032	.042	.027	**.021**	**.023**
	SI-$F_{\max}$ ↑	U	–	**.896**	**.869**	–	–	–	–	–	.864	.746	.822	.865	**.897**
		S	–	**.937**	.917	–	–	–	–	–	.918	.838	.897	**.926**	**.944**
		A	–	**.949**	.933	–	–	–	–	–	.934	.871	.918	**.941**	**.955**
ESDIs-SOD	$S_m \uparrow$	U	–	–	–	**.824**	.793	–	–	–	**.825**	.783	.767	.562	**.830**
		S	–	–	–	**.888**	.873	–	–	–	**.889**	.857	.851	.642	**.893**
		A	–	–	–	**.893**	**.882**	–	–	–	**.893**	.867	.859	.676	**.899**
	SI-MAE ↓	U	–	–	–	.083	.099	–	–	–	**.078**	**.077**	.086	.205	**.066**
		S	–	–	–	**.053**	.063	–	–	–	**.049**	**.049**	.055	.214	**.043**
		A	–	–	–	.058	.068	–	–	–	**.052**	**.053**	.059	.201	**.048**
	SI-$F_{\max}$ ↑	U	–	–	–	**.880**	.873	–	–	–	**.875**	.788	.835	.759	**.884**
		S	–	–	–	**.912**	**.916**	–	–	–	.908	.846	.885	.804	**.920**
		A	–	–	–	**.916**	**.921**	–	–	–	.913	.858	.891	.813	**.925**
ORSI-4199	$S_m \uparrow$	U	–	–	**.791**	–	–	**.793**	–	–	.790	**.793**	.789	.777	**.795**
		S	–	–	**.859**	–	–	**.860**	–	–	.855	.853	.845	.831	**.863**
		A	–	–	**.877**	–	–	**.878**	–	–	.875	.871	.852	.846	**.881**
	SI-MAE ↓	U	–	–	**.024**	–	–	**.024**	–	–	**.025**	.028	**.025**	.037	**.023**
		S	–	–	**.024**	–	–	**.025**	–	–	**.025**	**.025**	**.024**	.037	**.022**
		A	–	–	**.025**	–	–	**.026**	–	–	**.028**	**.028**	.033	.040	**.026**
	SI-$F_{\max}$ ↑	U	–	–	.857	–	–	**.870**	–	–	**.871**	.830	.845	.823	**.877**
		S	–	–	.902	–	–	**.914**	–	–	**.911**	.885	.894	.864	**.919**
		A	–	–	.920	–	–	**.929**	–	–	**.927**	.907	.907	.885	**.934**
SBU Shadow	$S_m \uparrow$	U	–	–	–	–	–	–	.812	**.832**	**.826**	.769	.815	.568	**.843**
		S	–	–	–	–	–	–	**.849**	**.868**	.846	.815	.843	.555	**.876**
		A	–	–	–	–	–	–	.876	**.896**	**.880**	.858	.875	.609	**.904**
	SI-MAE ↓	U	–	–	–	–	–	–	.068	**.039**	.047	.055	**.046**	.120	**.038**
		S	–	–	–	–	–	–	.103	**.078**	.094	.104	**.091**	.202	**.077**
		A	–	–	–	–	–	–	.095	**.069**	**.083**	.093	**.083**	.196	**.071**
	SI-$F_{\max}$ ↑	U	–	–	–	–	–	–	**.885**	.854	**.859**	.721	.818	.667	**.887**
		S	–	–	–	–	–	–	**.877**	.846	**.850**	.774	.831	.619	**.894**
		A	–	–	–	–	–	–	**.905**	.883	**.887**	.829	.868	.694	**.921**
GSD	$S_m \uparrow$	U	–	–	–	–	–	–	**.905**	**.907**	.850	**.905**	.812	.663	**.948**
		S	–	–	–	–	–	–	.880	**.905**	**.886**	.868	.847	.642	**.915**
		A	–	–	–	–	–	–	.848	**.885**	**.876**	.873	.846	.621	**.896**
	SI-MAE ↓	U	–	–	–	–	–	–	.031	**.019**	.042	.044	**.028**	.067	**.017**
		S	–	–	–	–	–	–	.069	**.049**	**.056**	.057	.068	.155	**.047**
		A	–	–	–	–	–	–	.084	**.054**	**.055**	.058	.066	.187	**.048**
	SI-$F_{\max}$ ↑	U	–	–	–	–	–	–	**.970**	**.969**	.963	.961	.961	.941	**.978**
		S	–	–	–	–	–	–	**.961**	.958	**.960**	.942	.943	.908	**.970**
		A	–	–	–	–	–	–	**.964**	**.961**	**.964**	.949	.947	.904	**.972**
PROMISE12	SI-MAE ↓	U	–	–	–	–	–	–	.023	**.007**	**.007**	**.008**	.012	.023	**.004**
		S	–	–	–	–	–	–	.017	**.005**	.008	.009	**.007**	.031	**.004**
		A †	–	–	–	–	–	–	.017	**.005**	.008	.009	**.007**	.031	**.004**
	SI-$F_{\max}$ ↑	U	–	–	–	–	–	–	.856	.859	**.887**	**.874**	.857	.552	**.895**
		S	–	–	–	–	–	–	**.934**	.922	**.947**	.933	.919	.564	**.948**
		A †	–	–	–	–	–	–	**.934**	.922	**.947**	.933	.919	.564	**.948**
	mIoU ↑	U	–	–	–	–	–	–	.620	.758	**.789**	**.791**	.707	.597	**.806**
		S	–	–	–	–	–	–	.833	.894	**.904**	**.905**	.877	.675	**.912**
		A †	–	–	–	–	–	–	.833	.894	**.904**	**.905**	.877	.675	**.912**
TN3K	SI-MAE ↓	U	–	–	–	–	–	–	.019	.028	**.012**	**.015**	.036	.082	**.007**
		S	–	–	–	–	–	–	**.024**	.030	**.018**	**.020**	.030	.063	**.018**
		A	–	–	–	–	–	–	.043	.037	**.021**	**.026**	.035	.085	**.025**
	SI-$F_{\max}$ ↑	U	–	–	–	–	–	–	**.897**	.877	**.895**	.828	.879	.849	**.910**
		S	–	–	–	–	–	–	**.917**	.897	**.924**	.869	.903	.867	**.929**
		A	–	–	–	–	–	–	**.935**	.908	**.943**	.892	.914	.870	**.939**
	mIoU ↑	U	–	–	–	–	–	–	**.740**	.736	**.793**	.729	.711	.592	**.803**
		S	–	–	–	–	–	–	**.823**	.810	**.845**	.816	.806	.668	**.846**
		A	–	–	–	–	–	–	.813	.830	**.861**	**.845**	.830	.658	**.864**
Best / Reported			0/9	0/18	1/18	0/9	0/9	0/9	0/36	1/36	3/72	0/72	0/72	3/72	**67/72**

Qualitative Results. Fig. 4 illustrates context-dependent foreground selection in RELIC, where different concept specifications produce distinct or overlapping responses for the same image. The owl is highlighted under both camouflage and saliency conditions, consistent with the non-exclusive relationship between these concepts discussed in previous ranking studies (Fang et al., 2021; Tian et al., 2022; Xu et al., 2026a), while the shadow condition selects the surrounding tree cavity. In the indoor scene, the transparent condition highlights the window, whereas saliency and camouflage conditions select the person. Similar response overlap appears across domains: for the remote-sensing scene, RELIC produces consistent responses when conditioned on saliency-related concepts from remote-sensing, industrial, and natural scenes. For the ultrasound image, thyroid-nodule and camouflage conditions highlight the nodule, while the shadow condition emphasizes darker surrounding regions. Together, these examples suggest that heterogeneous CD concepts can share contextual cues while requiring different foreground interpretations, illustrating how support-defined concepts guide local responses across visually distinct domains. Fig. 5 further evaluates the segmentation quality of RELIC against existing unified methods. Across the eight tasks, RELIC more consistently retains

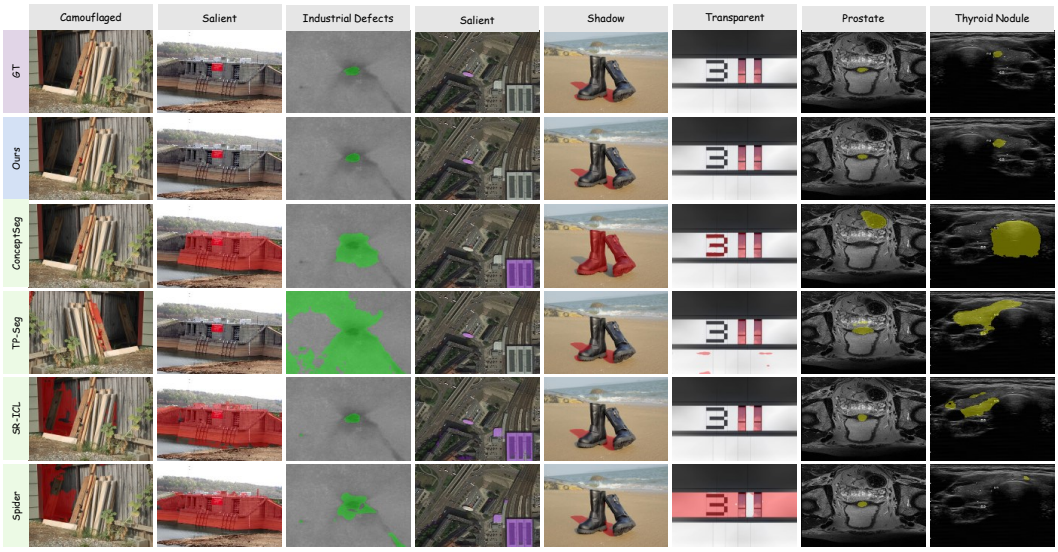

Figure 5: Visual comparison of RELIC with other unified methods.

Table 3: Ablation studies on the eight tasks. See § E for SI-$F_{\max}$ and SI-MAE results.

Variant	COD10K $S_m \uparrow$			DUTS $S_m \uparrow$			ESDIs-SOD $S_m \uparrow$			ORSI-4199 $S_m \uparrow$			SBU Shadow $S_m \uparrow$			GSD $S_m \uparrow$			PROMISE12 mIoU ↑			TN3K mIoU ↑		
	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A
(a) Components and conditioning positions																								
Tail-only	.692	.829	.852	.728	.870	.893	.745	.840	.851	.749	.824	.840	.793	.806	.850	.927	.865	.845	.774	.888	.888	.756	.818	.840
+ Representation Conditioning	.724	.841	.862	.821	.902	.912	.777	.859	.867	.770	.841	.858	.789	.844	.874	.884	.846	.832	.762	.880	.880	.746	.819	.838
Relative change (%)	+4.6	+1.4	+1.2	+12.8	+3.7	+2.1	+4.3	+2.3	+1.9	+2.8	+2.1	+2.1	-0.5	+4.7	+2.8	-4.6	-2.2	-1.5	-1.6	-0.9	-0.9	-1.3	+0.1	-0.2
+ SC-SEV	.746	.855	.876	.825	.913	.927	.801	.879	.885	.784	.856	.880	.827	.872	.899	.931	.904	.885	.777	.907	.907	.793	.846	.859
Relative change (%)	+7.8	+3.1	+2.8	+13.3	+4.9	+3.8	+7.5	+4.6	+4.0	+4.7	+3.9	+4.8	+4.3	+8.2	+5.8	+0.4	+4.5	+4.7	+0.4	+2.1	+2.1	+4.9	+3.4	+2.3
+ EGST-MoE (Full RELIC)	**.764**	**.879**	**.894**	**.847**	**.924**	**.935**	**.830**	**.893**	**.899**	**.795**	**.863**	**.881**	**.843**	**.876**	**.904**	**.948**	**.915**	**.896**	**.806**	**.912**	**.912**	**.803**	**.846**	**.864**
Relative change (%)	+10.4	+6.0	+4.9	+16.3	+6.2	+4.7	+11.4	+6.3	+5.6	+6.1	+4.7	+4.9	+6.3	+8.7	+6.4	+2.3	+5.8	+6.0	+4.1	+2.7	+2.7	+6.2	+3.4	+2.9
(b) SC-SEV																								
w/o Descriptor Group Matching	.710	.828	.848	.820	.899	.911	.776	.857	.866	.764	.839	.861	.808	.851	.878	.932	.888	.865	.731	.863	.863	.784	.827	.851
Relative change (%)	-4.8	-3.2	-3.2	-0.6	-1.5	-1.7	-3.1	-2.5	-2.1	-2.6	-2.0	-2.2	-2.3	-2.4	-2.3	+0.1	-1.8	-2.3	-5.9	-4.9	-4.9	-1.1	-2.2	-0.9
w/o Size-aware Verification	.714	.831	.851	.822	.901	.912	.781	.860	.868	.776	.842	.862	.811	.854	.880	.934	.888	.864	.744	.871	.871	.783	.828	.849
Relative change (%)	-4.3	-2.8	-2.9	-0.4	-1.3	-1.6	-2.5	-2.2	-1.9	-1.0	-1.6	-2.0	-1.9	-2.1	-2.1	+0.3	-1.8	-2.4	-4.2	-4.0	-4.0	-1.3	-2.1	-1.2
w/o Concept-aware Verification	.715	.832	.852	.821	.898	.910	.786	.861	.867	.768	.841	.862	.816	.854	.880	.915	.885	.864	.716	.858	.858	.775	.828	.854
Relative change (%)	-4.2	-2.7	-2.7	-0.5	-1.6	-1.8	-1.9	-2.0	-2.0	-2.0	-1.8	-2.0	-1.3	-2.1	-2.1	-1.7	-2.1	-2.4	-7.9	-5.4	-5.4	-2.3	-2.1	-0.6
Full SC-SEV	**.746**	**.855**	**.876**	**.825**	**.913**	**.927**	**.801**	**.879**	**.885**	**.784**	**.856**	**.880**	**.827**	**.872**	**.899**	**.931**	**.904**	**.885**	**.777**	**.907**	**.907**	**.793**	**.846**	**.859**
(c) Experts in EGST-MoE																								
w/o Weak-evidence Expert (E_w)	.732	.841	.862	.835	.916	.926	.805	.877	.885	.782	.855	.876	.836	.867	.895	.945	.882	.869	.774	.903	.903	.785	.842	.858
Relative change (%)	-4.2	-4.3	-3.6	-1.4	-0.9	-1.0	-3.0	-1.8	-1.6	-1.6	-0.9	-0.6	-0.8	-1.0	-1.0	-0.3	-3.6	-3.0	-4.0	-1.0	-1.0	-2.2	-0.5	-0.7
w/o Boundary Expert (E_b)	.728	.839	.861	.838	.917	.927	.805	.877	.885	.791	.859	.876	.833	.868	.894	.940	.883	.872	.788	.904	.904	.766	.839	.860
Relative change (%)	-4.7	-4.6	-3.7	-1.1	-0.8	-0.9	-3.0	-1.8	-1.6	-0.5	-0.5	-0.6	-1.2	-0.9	-1.1	-0.8	-3.5	-2.7	-2.2	-0.9	-0.9	-4.6	-0.8	-0.5
w/o Fragment Expert (E_f)	.728	.844	.864	.840	.918	.927	.801	.874	.882	.788	.857	.876	.834	.866	.893	.939	.886	.875	.786	.903	.903	.784	.848	.862
Relative change (%)	-4.7	-4.0	-3.4	-0.8	-0.6	-0.9	-3.5	-2.1	-1.9	-0.9	-0.7	-0.6	-1.1	-1.1	-1.2	-0.9	-3.2	-2.3	-2.5	-1.0	-1.0	-2.4	+0.2	-0.2
w/o FP-suppression Expert (E_{fp})	.738	.846	.865	.840	.919	.927	.810	.878	.885	.790	.859	.877	.834	.869	.895	.917	.883	.872	.787	.902	.902	.763	.844	.862
Relative change (%)	-3.4	-3.8	-3.2	-0.8	-0.5	-0.9	-2.4	-1.7	-1.6	-0.6	-0.5	-0.5	-1.1	-0.8	-1.0	-3.3	-3.5	-2.7	-2.4	-1.1	-1.1	-5.0	-0.2	-0.2
Full EGST-MoE	**.764**	**.879**	**.894**	**.847**	**.924**	**.935**	**.830**	**.893**	**.899**	**.795**	**.863**	**.881**	**.843**	**.876**	**.904**	**.948**	**.915**	**.896**	**.806**	**.912**	**.912**	**.803**	**.846**	**.864**
(d) Joint vs. separate training																								
Separate Training	.725	.842	.863	.817	.906	.922	.783	.870	.878	.773	.847	.868	.838	.874	.899	.946	.898	.882	.793	.906	.906	.773	.845	.863
Joint Training	**.764**	**.879**	**.894**	**.847**	**.924**	**.935**	**.830**	**.893**	**.899**	**.795**	**.863**	**.881**	**.843**	**.876**	**.904**	**.948**	**.915**	**.896**	**.806**	**.912**	**.912**	**.803**	**.846**	**.864**

small target regions, recovers incomplete structures, and reduces background responses. In contrast, competing methods more frequently miss weak targets or produce fragmented and excessive predictions. The visual results are consistent with the quantitative improvements and support the ability of RELIC to preserve and refine concept-relative evidence across heterogeneous CD concepts.

4.3 Ablation Study

Distributed Concept Conditioning. Tab. 3(a) progressively introduces representation conditioning, SC-SEV, and EGST-MoE. Representation conditioning alone yields mixed results, improving 15/24 settings while degrading the remaining nine. Adding SC-SEV improves all 24 settings over representation conditioning alone, with average relative gains of 3.35% (U) and 2.94% (S). The complete RELIC improves all 24 settings over Tail-only, achieving average relative gains of 7.90% (U) and 5.48% (S). These results support combining representation adaptation with explicit evidence verification and subsequent refinement.

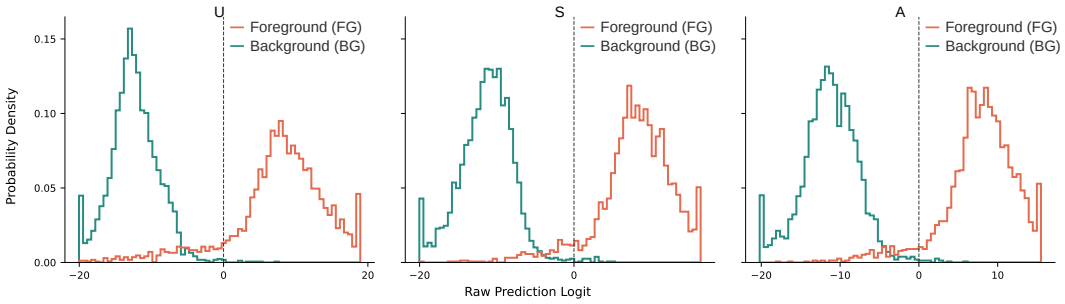

Figure 6: Foreground/background logit distributions across the U/S/A subsets. Dashed lines mark the zero-logit decision boundary.

Effectiveness of SC-SEV. Tab. 3(b) evaluates descriptor-group matching, size-aware verification, and concept-aware verification. Removing these components lowers scores in 70/72 comparisons, with average relative drops ranging from 1.91% to 2.72% on U and from 2.23% to 2.56% on S. Concept-aware verification is particularly important for ultra-small targets in PROMISE12, where its removal reduces mIoU from 0.777 to 0.716 (-7.9%). The overall results support the complementary contributions of support descriptor diversity and adaptive verification.

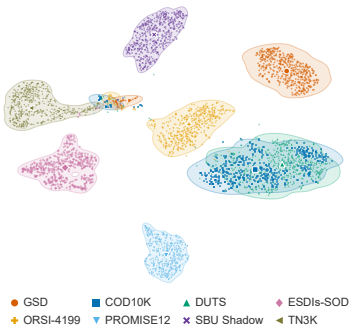

Figure 7: t-SNE visualization of task-conditioned representations after dynamic concept prediction.

Functional Experts in EGST-MoE. Tab. 3(c) individually removes the weak-evidence, boundary, fragment, and false-positive suppression experts. Scores decrease in 95/96 comparisons, including all 32 ultra-small comparisons. Across the four removals, average relative drops range from 2.10% to 2.37% on U and from 1.51% to 1.75% on S. The varying impact across datasets supports the complementary contributions of the heterogeneous experts to small-target refinement. Additional controlled ablations on expert heterogeneity, expert weighting, and spatial gating are provided in § E.

Joint Training vs. Separate Training. Tab. 3(d) compares joint and separate training using the same architecture and optimization settings. Joint training improves all 24 dataset–scale settings, with average relative gains of 3.03% (U), 1.73% (S), and 1.48% (A). These results support the benefits of jointly learning heterogeneous CD tasks within a shared, support-conditioned framework.

Prediction and Representation Analysis. Fig. 6 provides evidence of foreground–background discrimination across scales. Foreground logits are concentrated mainly above zero and background logits below zero, including on the U and S subsets, although their distribution tails overlap. This indicates that the predictions retain discriminative responses under limited spatial support. Fig. 7 visualizes the task-conditioned representations after dynamic concept prediction. Several datasets form distinct clusters, while COD10K and DUTS exhibit substantial overlap in the projection. This coexistence of task-associated structure and overlapping representations is consistent with the shared responses observed under camouflage and saliency conditions in Fig. 4.

5 Conclusion

We present RELIC, a unified framework for context-dependent concept segmentation through reliable small-target evidence learning. By integrating distributed concept conditioning, adaptive evidence verification, and state-guided refinement, RELIC assesses and refines sparse query responses according to support-defined concepts within a shared model. Experiments on eight benchmarks demonstrate consistently strong performance across natural, industrial, remote-sensing, and medical scenarios, with pronounced advantages on small and ultra-small targets. Our findings suggest that concept-relative evidence learning provides a useful perspective for unified segmentation when spatial support is limited.

AI USE STATEMENT

Generative AI was used to assist with methodological refinement, feedback on experimental design, code review and debugging, checks of data-processing and evaluation procedures, and interpretation of experimental results. It also supported manuscript drafting, language editing, and the preparation of figures and tables. The authors reviewed and revised the AI-assisted material and retained control over all research decisions and final conclusions. The authors take full responsibility for verifying the technical claims, code, results, and references, and for the accuracy, originality, and integrity of the final manuscript and associated artifacts, including compliance with academic standards, intellectual property rights, and applicable laws.

REFERENCES

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. UniverSeg: Universal medical image segmentation. In *ICCV*, pp. 21438–21451, 2023.

Shijie Chang, Xiaoqi Zhao, Lihe Zhang, and Tiancheng Wang. Unified medical lesion segmentation via self-referring indicator. In *CVPR*, pp. 10414–10424, 2025.

Wenqi Cui, Kechen Song, Hu Feng, Xiujian Jia, Shaoning Liu, and Yunhui Yan. Autocorrelation-aware aggregation network for salient object detection of strip steel surface defects. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2023.

Wei Dai, Rui Liu, Zixuan Wu, Tianyi Wu, Min Wang, Junxian Zhou, Yixuan Yuan, and Jun Liu. Exploiting scale-variant attention for segmenting small medical objects. *IEEE TNNLS*, 37(7): 3161–3175, 2026.

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pp. 4548–4557, 2017.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, pp. 2777–2787, 2020a.

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. PraNet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, pp. 263–273, 2020b.

Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic COVID-19 lung infection segmentation from CT images. *IEEE TMI*, 39(8):2626–2637, 2020c.

Hao Fang, Daoxin Zhang, Yi Zhang, Minghao Chen, Jiawei Li, Yao Hu, Deng Cai, and Xiaofei He. Salient object ranking with position-preserved attention. In *ICCV*, pp. 16331–16341, 2021.

Haifan Gong, Jiaxin Chen, Guanqi Chen, Haofeng Li, Guanbin Li, and Fei Chen. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Computers in Biology and Medicine*, 155:106389, 2023.

Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *CVPR*, pp. 22046–22055, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Lingjie Kong, Qiaoling Wei, Chengming Xu, Han Chen, and Yanwei Fu. EFCNet: Every feature counts for small medical object segmentation. *arXiv preprint arXiv:2406.18201*, 2024.

Feiran Li, Qianqian Xu, Shilong Bao, Zhiyong Yang, Runmin Cong, Xiaochun Cao, and Qingming Huang. Size-invariance matters: Rethinking metrics and losses for imbalanced multi-object salient object detection. In *ICML*, pp. 28989–29021, 2024.

Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, and Yu-Chiang Frank Wang. Recover and identify: A generative dual model for cross-resolution person re-identification. In *ICCV*, pp. 8090–8099, 2019.

Jiaying Lin, Zebang He, and Rynson WH Lau. Rich context aggregation with reflection prior for glass surface detection. In *CVPR*, pp. 13415–13424, 2021.

Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Medical Image Analysis*, 18(2):359–373, 2014.

Ange Lou, Shuyue Guan, Hanseok Ko, and Murray H Loew. CaraNet: Context axial reverse attention network for segmentation of small medical objects. In *Medical Imaging 2022: Image Processing*, volume 12032, pp. 81–92, 2022.

Lei Lu, Xinyu Xiong, Lianjun Liu, and Hao Tang. BG-UISNet: A novel boundary-guided network for robust and precise underwater image segmentation. *IEEE Sensors Journal*, 2025.

Kaiyi Luo, Bangjun Wang, Li Zhang, Fanzhang Li, Fei Zhu, and Jiaqing Fan. ACO-MoE-LoRA: Evolving-while-training for adapting segment anything model 2 to specialized domains. In *ICML*, 2026a.

Ziyang Luo, Nian Liu, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Shahbaz Khan, and Junwei Han. VSCode-v2: Dynamic prompt learning for general visual salient and camouflaged object detection with two-stage optimization. *IEEE TPAMI*, 48(3):3137–3153, 2026b.

Jianhua Ma, Mingfeng Jiang, Xian Fang, Jiatong Chen, Yaming Wang, and Guang Yang. Hybrid aggregation strategy with double inverted residual blocks for lightweight salient object detection. *Neural Networks*, 194:108097, 2026.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, pp. 565–571, 2016.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, pp. 2160–2170, 2022.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. ZoomNeXt: A unified collaborative pyramid network for camouflaged object detection. *IEEE TPAMI*, 46(12):9205–9220, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*, 2025.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

Kunye Shen, Xiaofei Zhou, and Zhi Liu. Class knowledge-guided lightweight network for salient object detection of strip steel surface defect. *IEEE TCSVT*, 2026.

Jing Sun, Siyu Liang, and Yu Zhao. Unleashing the potential of segment anything model 2 for efficient ORSI salient object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2026a.

Yanguang Sun, Chao Wang, Jian Yang, and Lei Luo. Small but mighty: Dynamic wavelet expert-guided fine-tuning of large-scale models for optical remote sensing object segmentation. In *AAAI*, pp. 9224–9232, 2026b.

Xin Tian, Ke Xu, Xin Yang, Lin Du, Baocai Yin, and Rynson WH Lau. Bi-directional object-context prioritization learning for saliency ranking. In *CVPR*, pp. 5882–5891, 2022.

Zhengzheng Tu, Chao Wang, Chenglong Li, Minghao Fan, Haifeng Zhao, and Bin Luo. ORSI salient object detection via multiscale joint region and boundary model. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *ECCV*, pp. 816–832, 2016.

Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pp. 136–145, 2017.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, pp. 6830–6839, 2023a.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Towards segmenting everything in context. In *ICCV*, pp. 1130–1140, 2023b.

Jun Wei, Shuhui Wang, and Qingming Huang. F^3Net: Fusion, feedback and focus for salient object detection. In *AAAI*, pp. 12321–12328, 2020.

Jiawei Xu, Qiangqiang Zhou, Zhouping Li, Yanjiao Shi, Yugen Yi, and Jiacong Yu. HVPNet: A bio-inspired network for general salient and camouflaged object detection. *Neural Networks*, pp. 109340, 2026a.

Jiawei Xu, Qiangqiang Zhou, Dandan Zhu, Yong Chen, Yugen Yi, and Xiaoqi Zhao. TP-Seg: Task-prototype framework for unified medical lesion segmentation. In *CVPR*, pp. 5452–5462, 2026b.

Chenhongyi Yang, Zehao Huang, and Naiyan Wang. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In *CVPR*, pp. 13668–13677, 2022.

Zhen Yang, Boyang Jiao, and Xiangyang Ren. RAC-Net: Interpretable medical small target segmentation network with x-ray radiation attenuation characterization. *IEEE TPAMI*, 48(10):12975–12994, 2026.

Sheng Ye, Xin Chen, Yan Zhang, Xianming Lin, and Liujuan Cao. ESCNet: Edge-semantic collaborative network for camouflaged object detection. In *ICCV*, pp. 20053–20063, 2025.

Xiaoqi Zhao, Youwei Pang, Shijie Chang, Yuan Zhao, Lihe Zhang, Chenyang Yu, Hanqi Liu, Jiaming Zuo, Jinsong Ouyang, Weisi Lin, Georges El Fakhri, Huchuan Lu, and Xiaofeng Liu. Inspiring the next generation of segment anything models: Comprehensively evaluate SAM and SAM 2 with diverse prompts towards context-dependent concepts under different scenes. *arXiv preprint arXiv:2412.01240*, 2024a.

Xiaoqi Zhao, Youwei Pang, Wei Ji, Baicheng Sheng, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Spider: A unified framework for context-dependent concept segmentation. In *ICML*, pp. 60906–60926, 2024b.

Yuan Zhao, Youwei Pang, Jiaming Zuo, Wei Ji, Kailai Zhou, Bin Fan, Yunkang Cao, Lihe Zhang, Xiaofeng Liu, Huchuan Lu, Weisi Lin, Dacheng Tao, and Xiaoqi Zhao. ConceptSeg-R1: Segment any concept via meta-reinforcement learning. *arXiv preprint arXiv:2605.20385*, 2026.

Qiangqiang Zhou, Jiawei Xu, Yong Chen, Dandan Zhu, Yugen Yi, and Xiaoqi Zhao. DifferSeg: Towards diverse multimodal binary segmentation via differential perception and frequency guidance. *IEEE TCSVT*, 2026.

Haoran Zhu, Wen Yang, Guangyou Yang, Chang Xu, Ruixiang Zhang, Fang Xu, Haijian Zhang, and Gui-Song Xia. Generalized small object detection: A point-prompted paradigm and benchmark. *arXiv preprint arXiv:2604.02773*, 2026.

Yabin Zhu, Chenglong Li, Yao Liu, Xiao Wang, Jin Tang, Bin Luo, and Zhixiang Huang. Tiny object tracking: A large-scale dataset and a baseline. *IEEE TNNLS*, 35(8):10273–10287, 2024.

A Comparison Configuration

Following recent unified segmentation frameworks, adopting strong pretrained backbones to adapt to diverse visual concepts has become a common practice (Zhao et al., 2024b; Chang et al., 2025; Xu et al., 2026b; Zhao et al., 2026). For fairness, all unified baselines for which we control the image preprocessing pipeline are evaluated at 512×512, unless their released implementations rely on a specific native resolution.

The compared unified methods employ strong pretrained visual backbones. Spider is evaluated with its ConvNeXt-L variant (Zhao et al., 2024b), and SR-ICL also adopts ConvNeXt-L (Chang et al., 2025). TP-Seg employs the Hiera-L encoder initialized from SAM 2 (Xu et al., 2026b), while RELIC similarly uses a frozen SAM 2 Hiera-L backbone. For ConceptSeg-R1, we use the publicly released ConceptSeg-R1-7B weights and retain its released input configuration. Its frozen SAM 3 segmentation branch operates at the native 1008×1008 resolution, while the multimodal reasoning branch follows the released preprocessing setting with a maximum pixel budget of 360,000 for each input image. Thus, the segmentation branch of ConceptSeg-R1 operates at a higher input resolution than the other unified methods evaluated at 512×512.

B Additional EGST-MoE Details

This section specifies the evidence statistics and spatial gates used by EGST-MoE. Both are computed from support-conditioned query evidence without using query ground truth.

Evidence State. Let $C_{\mathrm{rel},l} = \sigma(G_l/\tau_{\mathrm{rel},l})$ and $C_{\mathrm{abs},l} = \sigma((A_{\mathrm{fg},l} - \theta_l)/\tau_{\mathrm{abs},l})$ denote the relative and absolute verification factors in Eq. (3), with $V_l = C_{\mathrm{rel},l} \odot C_{\mathrm{abs},l}$. EGST-MoE summarizes these responses as $e_l = [o_l, p_l, v_l, h_l, f_l, m_l, r_{c,l}] \in \mathbb{R}^7$, with the first six entries computed as:

$$\begin{aligned} o_l &= M(V_l), & p_l &= \mathrm{TopMean}_{k_l}(V_l), \\ v_l &= \tfrac{1}{2}\big(M(|\nabla_x V_l|) + M(|\nabla_y V_l|)\big), & h_l &= M\big(4V_l \odot (1 - V_l)\big), \\ f_l &= M\big(|V_l - \mathrm{AvgPool}_3(V_l)|\big), & m_l &= M(G_l), \end{aligned} \tag{9}$$

where $M(\cdot)$ denotes spatial averaging, $\mathrm{TopMean}_{k_l}$ averages the largest $k_l = \min(4, H_l W_l)$ responses, and ∇_x, ∇_y denote local horizontal and vertical differences. The local-inconsistency term f_l provides a fragmentation cue, while $r_{c,l}$ follows the concept-agreement definition in SC-SEV. The router combines these statistics with the feature summaries u_l to predict expert weights in Eq. (4).

Implementation of Spatial Gating. We define $\mathrm{Close}_5(V_l) = \mathrm{MinPool}_5(\mathrm{MaxPool}_5(V_l))$ and compute the expert-specific gates from SC-SEV responses as:

$$\begin{aligned} \Gamma_{w,l} &= \big[C_{\mathrm{abs},l} \odot (1 - C_{\mathrm{rel},l})\big]_0^1, \quad \Gamma_{b,l} = \big[\mathrm{MaxPool}_3(V_l) - \mathrm{MinPool}_3(V_l)\big]_0^1, \\ \Gamma_{f,l} &= \Big[\big[\mathrm{Close}_5(V_l) - V_l\big]_+ + \mathrm{AvgPool}_5(V_l) \odot (1 - V_l)\Big]_0^1, \\ \Gamma_{fp,l} &= \big[\sigma(A_{\mathrm{fg},l}/\tau_{fp}) \odot \sigma(-G_l/\tau_{fp})\big]_0^1, \end{aligned} \tag{10}$$

where $[\cdot]_0^1$ clips values to $[0, 1]$, $[\cdot]_+ = \max(\cdot, 0)$, and $\tau_{fp} = 0.10$. Pooling subscripts denote the spatial window size, with unit stride and spatial resolution preserved. The weak-response and boundary gates emphasize uncertain foreground evidence and local transitions, while the fragment and false-positive gates focus on incomplete responses and foreground candidates contradicted by background evidence, respectively. For the false-positive expert, the attenuation map is $B_{fp,l} = \rho_{fp}\sigma(Q_{fp}(X_l)) \odot \Gamma_{fp,l}$, where $Q_{fp}(\cdot)$ denotes the expert output before sigmoid and $\rho_{fp} = 0.5\sigma(a_{fp})$ controls the suppression strength. The bounded map is broadcast across feature channels to form the suppression residual in Eq. (6).

C Object-Balanced Evidence Margin Loss

Small targets can be under-emphasized when foreground supervision is dominated by larger regions. Motivated by size-invariant object balancing (Li et al., 2024), we supervise the SC-SEV margin G_l

Table 4: SI-$F_{\max}$ ablations across eight datasets. Notation follows Table 3.

Variant	COD10K SI-$F_{\max}$ ↑			DUTS SI-$F_{\max}$ ↑			ESDIs-SOD SI-$F_{\max}$ ↑			ORSI-4199 SI-$F_{\max}$ ↑			SBU Shadow SI-$F_{\max}$ ↑			GSD SI-$F_{\max}$ ↑			PROMISE12 SI-$F_{\max}$ ↑			TN3K SI-$F_{\max}$ ↑		
	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A
(a) Overall framework and distributed concept conditioning																								
Tail-only	.741	.818	.844	.853	.918	.935	.842	.882	.889	.838	.882	.897	.867	.856	.885	.942	.934	.937	.880	.928	.928	.894	.914	.924
+ Representation Conditioning	.733	.812	.839	.865	.923	.937	.855	.891	.897	.849	.892	.909	.843	.867	.893	.948	.938	.941	.887	.928	.928	.890	.913	.923
+ SC-SEV	.744	.820	.846	.872	.930	.945	.877	.914	.920	.869	.915	.934	.865	.885	.913	.971	.965	.968	.899	.947	.947	.903	.923	.934
+ EGST-MoE (Full RELIC)	**.776**	**.855**	**.876**	**.897**	**.944**	**.955**	**.884**	**.920**	**.925**	**.877**	**.919**	**.934**	**.887**	**.894**	**.921**	**.978**	**.970**	**.972**	**.895**	**.948**	**.948**	**.910**	**.929**	**.939**
Average Improvement	.036	.038	.032	.033	.019	.015	.025	.023	.022	.024	.022	.020	.028	.024	.023	.023	.024	.022	.005	.013	.013	.013	.011	.011
(b) SC-SEV component analysis																								
w/o Descriptor Group Matching	.719	.802	.831	.857	.914	.927	.853	.897	.906	.844	.900	.921	.834	.859	.893	.946	.948	.954	.868	.931	.931	.890	.918	.930
w/o Size-aware Verification	.721	.803	.833	.859	.915	.929	.853	.899	.908	.848	.901	.920	.844	.867	.899	.947	.949	.955	.876	.934	.934	.894	.918	.930
w/o Concept-aware Verification	.720	.804	.833	.860	.916	.929	.854	.899	.908	.848	.902	.923	.849	.873	.902	.946	.950	.956	.873	.932	.932	.888	.918	.931
Full SC-SEV	**.744**	**.820**	**.846**	**.872**	**.930**	**.945**	**.877**	**.914**	**.920**	**.869**	**.915**	**.934**	**.865**	**.885**	**.913**	**.971**	**.965**	**.968**	**.899**	**.947**	**.947**	**.903**	**.923**	**.934**
Average Improvement	.023	.016	.013	.013	.015	.016	.023	.015	.012	.022	.014	.012	.022	.018	.015	.024	.016	.013	.026	.014	.014	.012	.005	.003
(c) Functional expert analysis of EGST-MoE																								
w/o Weak-Evidence Expert (E_w)	.738	.813	.840	.866	.925	.940	.874	.913	.919	.866	.914	.934	.866	.890	.917	.971	.964	.968	.886	.935	.935	.885	.915	.925
w/o Boundary Expert (E_b)	.737	.812	.840	.867	.925	.940	.875	.912	.919	.866	.912	.931	.864	.889	.916	.971	.963	.968	.893	.935	.935	.882	.916	.926
w/o Fragment Expert (E_f)	.739	.815	.842	.870	.928	.942	.877	.915	.921	.868	.915	.934	.870	.892	.919	.972	.965	.971	.891	.936	.936	.885	.917	.926
w/o FP-Suppression Expert (E_{fp})	.742	.815	.842	.869	.928	.942	.878	.916	.922	.870	.917	.935	.869	.892	.919	.973	.966	.971	.890	.937	.937	.885	.919	.929
Full EGST-MoE	**.776**	**.855**	**.876**	**.897**	**.944**	**.955**	**.884**	**.920**	**.925**	**.877**	**.919**	**.934**	**.887**	**.894**	**.921**	**.978**	**.970**	**.972**	**.895**	**.948**	**.948**	**.910**	**.929**	**.939**
Average Improvement	.036	.040	.034	.028	.016	.013	.007	.005	.004	.009	.003	.000	.019	.002	.002	.005	.005	.002	.004	.011	.011	.025	.011	.012
(d) Joint Training vs. Separate Training																								
Separate Training	.746	.818	.844	.870	.925	.941	.881	.914	.920	.868	.913	.930	.868	.892	.919	.971	.965	.970	.894	.947	.947	.909	.927	.935
Joint Training	.776	.855	.876	.897	.944	.955	.884	.920	.925	.877	.919	.934	.887	.894	.921	.978	.970	.972	.895	.948	.948	.910	.929	.939

at the object level, giving each foreground object equal weight regardless of its area. For background supervision, we exclude locations near foreground boundaries and retain the 10% of eligible background positions with the largest margins as hard negatives. The loss is defined as:

$$\mathcal{L}_{\text{rank}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{4}\left[\text{ObjAvg}\big([\mu_{\text{fg}} - G_l^t]_+\big) + \text{HardBgAvg}\big([\mu_{\text{bg}} + G_l^t]_+\big)\right], \tag{11}$$

where G_l^t is the SC-SEV margin for task t at scale l, $[\cdot]_+ = \max(\cdot, 0)$, and $\mu_{\text{fg}} = \mu_{\text{bg}} = 0.2$. The operator ObjAvg averages the foreground loss within each object and then equally across objects, while HardBgAvg averages the selected hard-background locations. The loss encourages positive foreground and negative hard-background margins across all four scales. Query ground truth defines the regions for this training loss and is not required at inference.

D Routing-Balancing Objectives

RELIC contains a sparse Prompt MoE in the encoder and a dense EGST-MoE for evidence refinement. For the Prompt MoE, $\mathcal{L}_{\text{enc-bal}}$ follows the standard expert-balancing objective of sparsely gated MoE (Shazeer et al., 2017), which regularizes expert importance and routing load to avoid persistent expert preference. For EGST-MoE, all four experts remain active and are combined using dense query-level weights. We therefore balance their accumulated routing importance rather than discrete expert load. For a query mini-batch $\mathcal{B}$, the balancing term is computed as:

$$\bar{\pi}_e = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\pi_{i,e}, \qquad \mathcal{L}_{\text{ref-bal}} = \text{CV}^2\left([\bar{\pi}_1, \ldots, \bar{\pi}_E]\right), \tag{12}$$

where $\pi_{i,e}$ is the query-level routing weight, $\bar{\pi}_e$ is its batch mean, and $E = 4$. The squared coefficient of variation is $\text{CV}^2(\mathbf{x}) = \text{Var}(\mathbf{x})/(\text{Mean}(\mathbf{x})^2 + \epsilon)$, with $\epsilon > 0$ for numerical stability. This regularizer discourages routing collapse while preserving query-dependent expert weights.

E Additional Ablation Studies

Tab. 4 and Tab. 5 report the corresponding SI-$F_{\max}$ and SI-MAE ablations across all eight datasets, following the configurations in Tab. 3.

Controlled EGST-MoE Design Ablations. Tab. 6 further isolates three design choices in EGST-MoE. Homogeneous expert cores remove functional specialization, uniform expert weighting removes query-dependent expert weighting, and the variant without spatial gating applies expert corrections without location-specific modulation. The full design consistently performs better, supporting the complementary roles of expert heterogeneity, adaptive weighting, and spatially localized refinement.

Table 5: SI-MAE ablations across eight datasets (lower is better). Notation follows Table 3.

Variant	COD10K SI-MAE ↓			DUTS SI-MAE ↓			ESDIs-SOD SI-MAE ↓			ORSI-4199 SI-MAE ↓			SBU Shadow SI-MAE ↓			GSD SI-MAE ↓			PROMISE12 SI-MAE ↓			TN3K SI-MAE ↓		
	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A
(a) Overall framework and distributed concept conditioning																								
Tail-only	.073	.068	.066	.057	.041	.042	.082	.054	.060	.028	.028	.037	.053	.098	.087	.027	.069	.075	.011	.005	.005	.012	.023	.031
+ Representation Conditioning	.060	.061	.060	.015	.027	.032	.087	.056	.062	.027	.026	.032	.061	.087	.079	.031	.080	.086	.008	.007	.007	.014	.021	.030
+ SC-SEV	.059	.061	.059	.019	.025	.027	.082	.053	.059	.026	.024	.027	.050	.085	.077	.024	.057	.061	.006	.005	.005	.008	.018	.028
+ EGST-MoE (Full RELIC)	**.054**	**.050**	**.049**	**.015**	**.021**	**.023**	**.066**	**.043**	**.048**	**.023**	**.022**	**.026**	**.038**	**.077**	**.071**	**.017**	**.047**	**.048**	**.004**	**.004**	**.004**	**.007**	**.018**	**.025**
Average Improvement	.010	.013	.013	.015	.010	.011	.018	.011	.012	.004	.004	.006	.017	.013	.010	.010	.022	.026	.004	.002	.002	.004	.003	.004
(b) SC-SEV component analysis																								
w/o Descriptor Group Matching	.063	.066	.069	.016	.028	.036	.086	.059	.069	.031	.030	.038	.054	.093	.090	.027	.062	.070	.011	.012	.012	.010	.023	.038
w/o Size-aware Verification	.063	.066	.068	.017	.029	.036	.086	.058	.068	.028	.029	.038	.053	.092	.089	.028	.064	.073	.009	.011	.011	.010	.022	.038
w/o Concept-aware Verification	.062	.065	.068	.017	.030	.036	.085	.059	.069	.030	.029	.038	.052	.091	.088	.028	.065	.072	.014	.014	.014	.010	.023	.037
Full SC-SEV	**.059**	**.061**	**.059**	**.019**	**.025**	**.027**	**.082**	**.053**	**.059**	**.026**	**.024**	**.027**	**.050**	**.085**	**.077**	**.024**	**.057**	**.061**	**.006**	**.005**	**.005**	**.008**	**.018**	**.028**
Average Improvement	.003	.004	.009	-.002	.004	.009	.003	.006	.010	.003	.005	.010	.002	.007	.011	.003	.006	.010	.005	.007	.007	.002	.004	.010
(c) Functional expert analysis of EGST-MoE																								
w/o Weak-Evidence Expert (E_w)	.053	.058	.061	.021	.026	.026	.084	.053	.063	.028	.025	.028	.043	.082	.079	.017	.063	.066	.008	.012	.012	.005	.019	.032
w/o Boundary Expert (E_b)	.054	.058	.061	.020	.025	.025	.080	.052	.062	.026	.024	.029	.044	.083	.079	.018	.062	.064	.008	.012	.012	.009	.019	.031
w/o Fragment Expert (E_f)	.054	.057	.060	.019	.025	.025	.080	.052	.061	.025	.024	.029	.042	.081	.077	.021	.063	.066	.008	.012	.012	.007	.017	.031
w/o FP-Suppression Expert (E_{fp})	.053	.057	.061	.019	.025	.026	.082	.053	.063	.025	.024	.028	.043	.081	.078	.023	.061	.064	.007	.012	.012	.013	.019	.031
Full EGST-MoE	**.054**	**.050**	**.049**	**.015**	**.021**	**.023**	**.066**	**.043**	**.048**	**.023**	**.022**	**.026**	**.038**	**.077**	**.071**	**.017**	**.047**	**.048**	**.004**	**.004**	**.004**	**.007**	**.018**	**.025**
Average Improvement	.000	.008	.012	.005	.004	.003	.015	.009	.014	.003	.002	.003	.005	.005	.007	.003	.015	.017	.004	.008	.008	.002	.000	.006
(d) Joint Training vs. Separate Training																								
Separate Training	.062	.061	.060	.019	.028	.029	.074	.049	.054	.028	.025	.030	.043	.081	.074	.027	.059	.061	.004	.005	.005	.011	.018	.028
Joint Training	.054	.050	.049	.015	.021	.023	.066	.043	.048	.023	.022	.026	.038	.077	.071	.017	.047	.048	.004	.004	.004	.007	.018	.025

Table 6: Controlled design ablations of EGST-MoE using S_m and mIoU. Notation follows Table 3.

Variant	COD10K S_m ↑			DUTS S_m ↑			ESDIs-SOD S_m ↑			ORSI-4199 S_m ↑			SBU Shadow S_m ↑			GSD S_m ↑			PROMISE12 mIoU ↑			TN3K mIoU ↑		
	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A	U	S	A
EGST-MoE controlled design ablations																								
Homogeneous Expert Cores	.717	.843	.865	.828	.905	.915	.790	.864	.871	.782	.848	.865	.826	.863	.887	.932	.886	.864	.768	.893	.893	.763	.828	.844
Uniform Expert Weighting	.720	.842	.864	.822	.903	.914	.781	.863	.869	.777	.845	.862	.831	.867	.889	.931	.888	.863	.783	.893	.893	.768	.832	.845
w/o Spatial Gating	.712	.836	.859	.815	.901	.913	.768	.858	.866	.778	.849	.865	.829	.860	.885	.937	.879	.866	.785	.893	.893	.783	.835	.846
Full EGST-MoE	**.764**	**.879**	**.894**	**.847**	**.924**	**.935**	**.830**	**.893**	**.899**	**.795**	**.863**	**.881**	**.843**	**.876**	**.904**	**.948**	**.915**	**.896**	**.806**	**.912**	**.912**	**.803**	**.846**	**.864**

F Additional Representation Analysis

Cross-task Foreground/Background Separation. Fig. 8 extends the scale-conditioned analysis in the main paper to all eight CD tasks. Despite substantial domain variation, the final task-conditioned representations maintain clear foreground-background discrimination, which is further reflected by the final prediction logits.

G Scale-wise Evaluation Protocol

Consistent with prior small-object evaluation protocols (Dai et al., 2026), we define ultra-small and small targets using foreground-to-image area ratios of 1% and 10%, respectively. Following the object-wise evaluation protocol (Li et al., 2024), we treat foreground connected components in binary GT masks as object proxies and use 4-connectivity, consistent with the released implementation. For an image I_i of size $H_i \times W_i$, let $\{C_{i,j}\}_{j=1}^{K_i}$ denote its K_i GT foreground components. Its scale is determined by the largest-component area ratio, computed as:

$$r_i = \frac{\max_j |C_{i,j}|}{H_i W_i}, \tag{13}$$

where $|C_{i,j}|$ denotes the pixel count of component $C_{i,j}$. The three evaluation subsets are defined as:

$$\mathcal{D}_U = \{I_i : r_i < 0.01\}, \qquad \mathcal{D}_S = \{I_i : r_i < 0.10\}, \qquad \mathcal{D}_A = \mathcal{D}_{\text{test}}, \tag{14}$$

where $\mathcal{D}_{\text{test}}$ denotes the complete test set and $\mathcal{D}_U \subset \mathcal{D}_S \subset \mathcal{D}_A$. For S_m and mIoU, metrics are computed on the complete prediction and GT masks of each image and averaged within the corresponding subset. SI-MAE and SI-$F_{\max}$ follow the object-wise evaluation principle (Li et al., 2024): GT connected components define the object-level regions, and the resulting measurements are aggregated across objects and images.

H Threshold-wise Performance Analysis

We further compare RELIC with four unified methods using precision–recall and F-measure curves under U, S, and A. Fig. 9 reports the curves macro-averaged over the eight benchmarks. RELIC

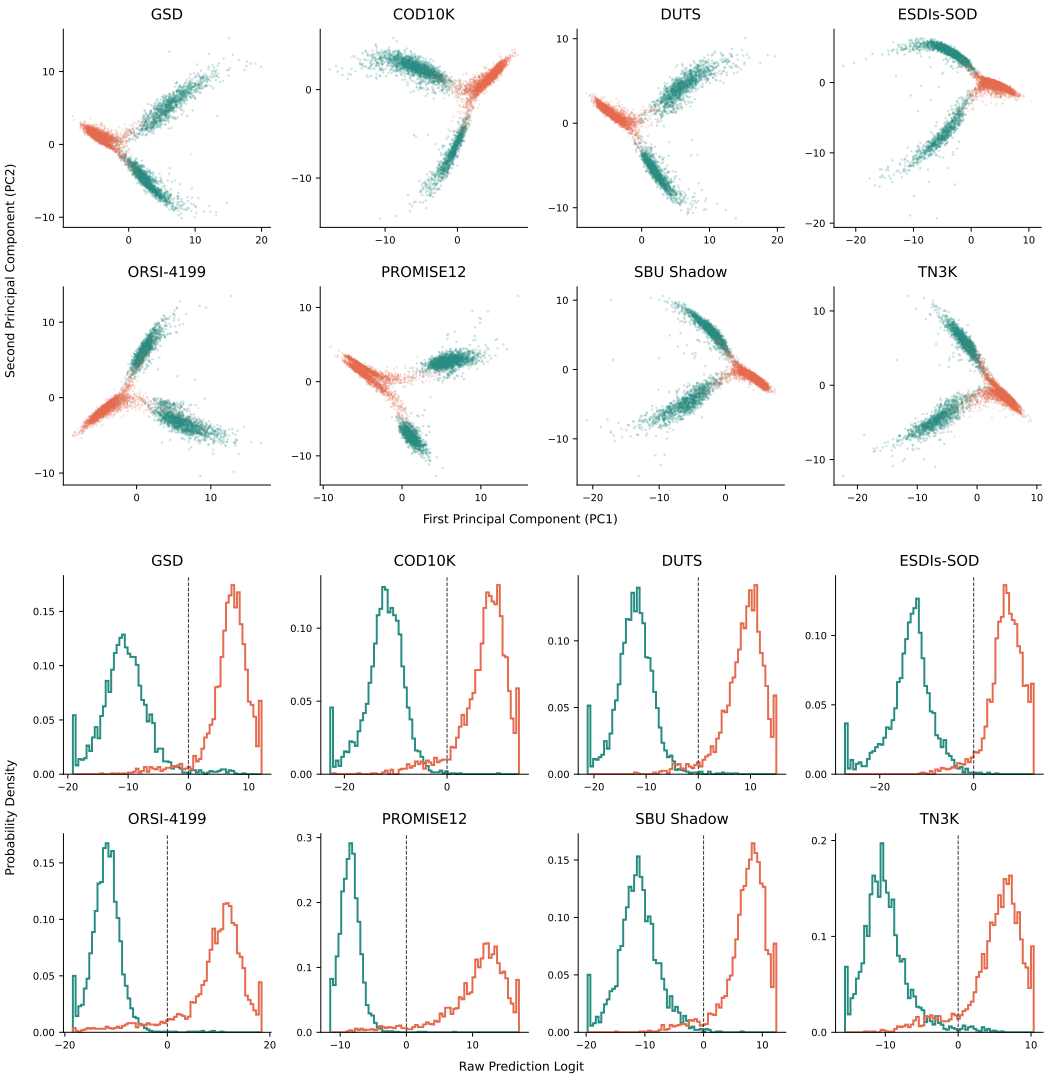

Figure 8: Foreground/background separation across eight CD concept segmentation tasks in the final feature and logit spaces.

reaches peak macro-averaged F-measures of 80.64%, 88.50%, and 90.41% on U, S, and A, respectively. Compared with Spider, the strongest unified baseline in this comparison, the margins are 5.16, 2.38, and 1.20 percentage points. The larger difference on U shows that the advantage of RELIC is most evident on ultra-small targets. The following figures provide the corresponding benchmark-wise comparisons. The value n in each panel denotes the number of images in the evaluation subset.

Ultra-small Targets. Fig. 10 presents the results on U. RELIC achieves the highest peak F-measure on six of the eight benchmarks, showing that the macro-level advantage is supported across multiple concepts rather than by a single benchmark. ConceptSeg performs best on COD10K, while Spider leads on PROMISE12.

Small Targets. Fig. 11 presents the results on S. The differences among the leading methods become smaller than on U, while RELIC maintains strong performance across the eight benchmarks. ConceptSeg performs particularly well on COD10K, and Spider remains competitive on PROMISE12.

All Targets. Fig. 12 reports the results on the complete test sets. The performance gap among the strongest unified methods is further reduced, while RELIC remains competitive across all eight benchmarks. Together with Fig. 9, the results show that the relative advantage of RELIC is more pronounced on ultra-small targets while remaining strong on the complete test sets.

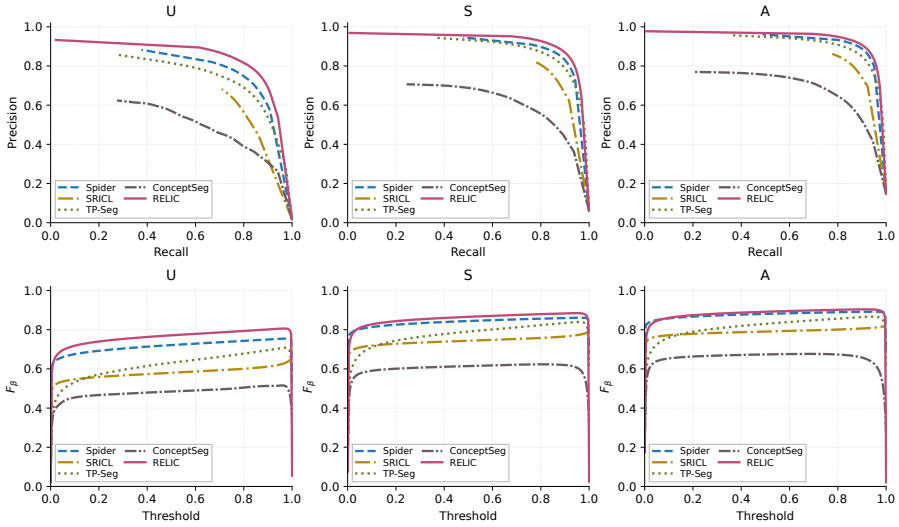

Figure 9: Macro-averaged precision–recall and F-measure curves of unified methods across eight benchmarks under U, S, and A.

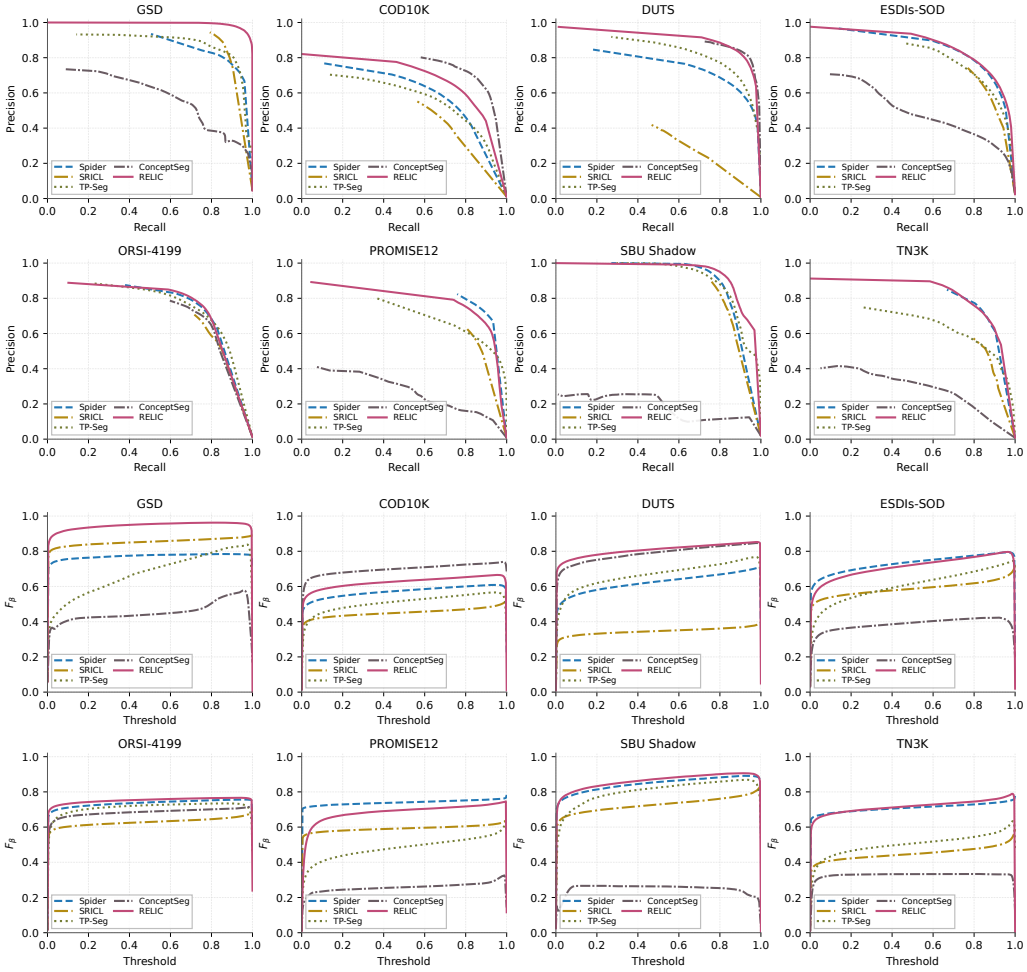

Figure 10: Benchmark-wise precision–recall and F-measure curves of unified methods on the ultra-small (U) subsets.

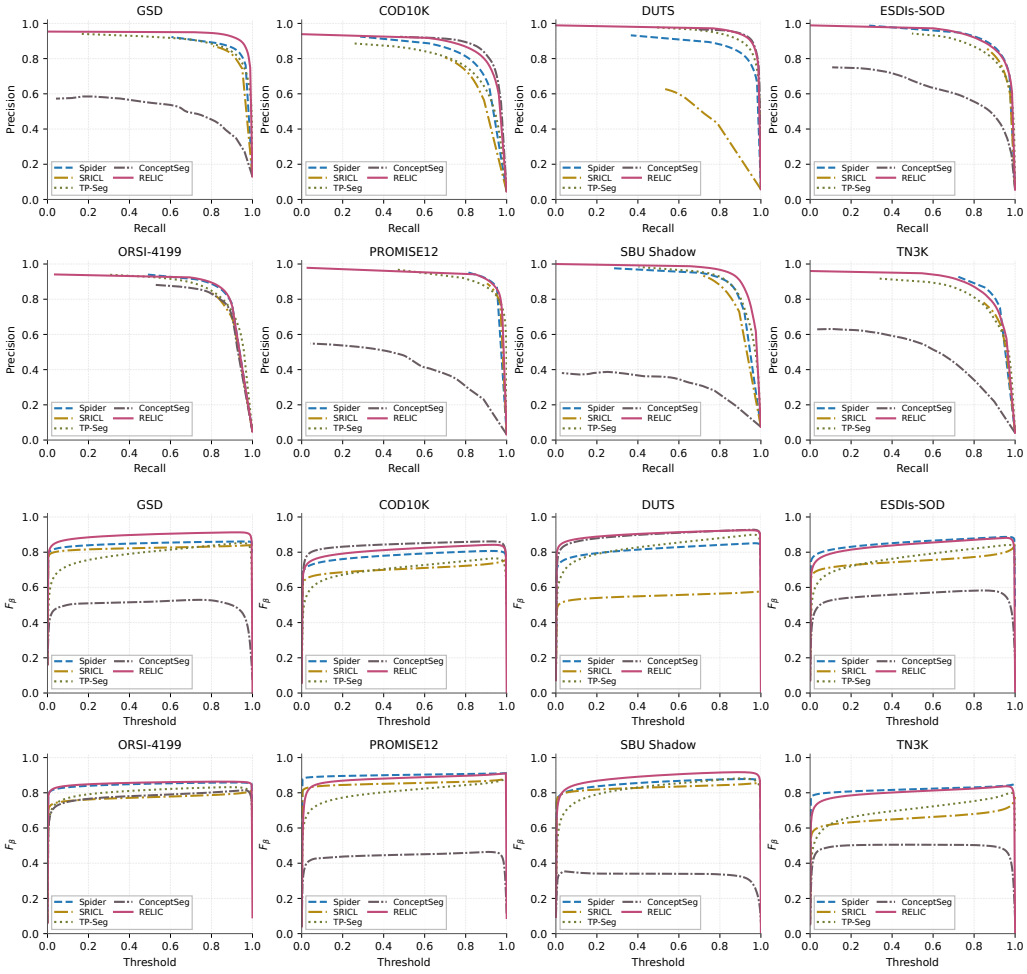

Figure 11: Benchmark-wise precision–recall and F-measure curves of unified methods on the small (S) subsets.

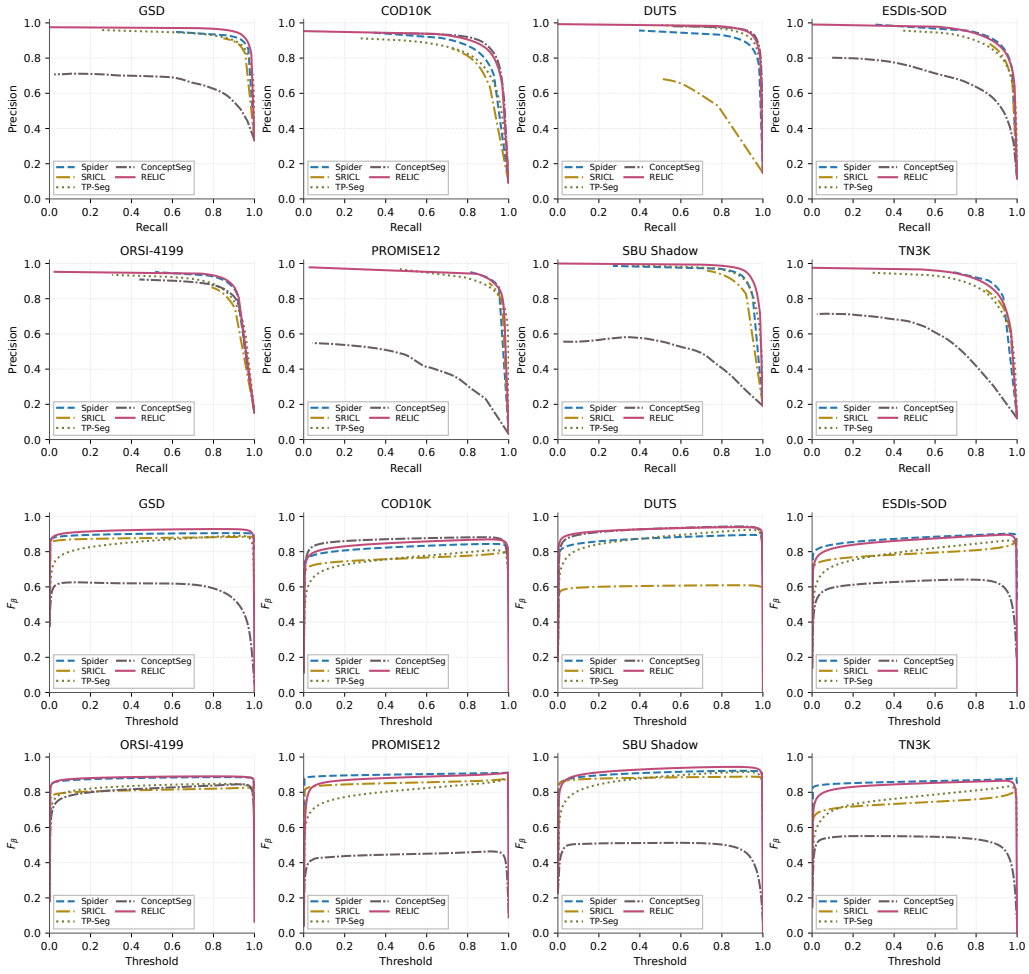

Figure 12: Benchmark-wise precision–recall and F-measure curves of unified methods on the complete (A) test sets.