EVIDENTIAL CURRICULUM LEARNING FOR MULTI-PROMPT WEAKLY SUPERVISED RGB-T SALIENT OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-prompt weakly supervised RGB-Thermal salient object detection (WSSOD) reduces annotation costs with points, boxes, and scribbles, but suffers from unreliable predictions and unstable training caused by incomplete prompts, cross-modal inconsistency, and imbalanced sample difficulty. To address these issues, we propose an Evidential Curriculum Learning Network (ECL-Net), which builds a unified uncertainty-driven learning pipeline from region-level reliability modeling to sample-level curriculum optimization. Specifically, Evidential Certain-Uncertain Refinement (ECUR) mechanism models Dirichlet evidence to decouple confident and uncertain regions and refines ambiguous features with certainty-weighted prototypes and intra-class self-attention. Evidential Uncertainty Curriculum Module (EUCM) reuses this uncertainty estimate for easy-to-hard sample scheduling, enabling stable RGB-T saliency priors before harder cases. Extensive experiments under point-, box-, and scribble-supervised settings on VT821, VT1000, and VT5000 demonstrate that ECL-Net achieves state-of-the-art performance against existing RGB-T WSSOD methods.

1 INTRODUCTION

RGB-Thermal salient object detection (RGB-T SOD) exploits complementary visible and thermal cues for robust perception under low illumination, clutter, and nighttime conditions, but dense mask annotation remains costly Li et al. (2021). Multi-prompt weak supervision with points, boxes, or scribbles reduces labeling effort yet provides only partial spatial evidence, making reliable cross-modal saliency learning a central challenge for WSSOD Wang et al. (2026).

Incomplete prompts make within-image prediction reliability highly heterogeneous, where confident, ambiguous, and unlabeled regions coexist. Treating all locations as equally reliable supervision solidifies early errors into pseudo-labels and lets uncertain features contaminate confident representations during cross-modal fusion Mai et al. (2026). Region-level unreliability thus requires explicit decoupling of certain and uncertain predictions.

Weakly annotated samples also differ in prompt quality and coverage across point, box, and scribble forms Hou et al. (2025). Exposing all samples equally from the start injects high-uncertainty gradients before stable saliency priors form, destabilizing optimization. Sample-level difficulty therefore calls for uncertainty-aware easy-to-hard curriculum scheduling.

To address these challenges, we propose ECL-Net, which couples region-level evidential refinement and sample-level curriculum scheduling through a shared uncertainty estimate. ECUR models foreground/background evidence with a Dirichlet distribution, decouples confident foreground, confident background, and uncertain regions, and refines uncertain features via certainty-weighted prototypes and intra-class self-attention within confident regions. EUCM reuses ECUR outputs to aggregate pixel-level ignorance and conflict uncertainties into sample-level difficulty scores and progressively admits low-difficulty samples from the unselected pool, enabling stable RGB-T saliency priors to be learned before adapting to ambiguous cases. Our contributions are summarized as follows:

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.

- To address unreliable predictions and uncertain-to-certain feature contamination caused by incomplete prompt supervision in WSSOD, we propose ECUR.
- To address sample difficulty imbalance induced by multi-prompt weak supervision with points, boxes, and scribbles, we propose EUCM.
- Extensive experiments under point-, box-, and scribble-supervised settings on VT821, VT1000, and VT5000 demonstrate that ECL-Net achieves state-of-the-art performance against existing WSSOD methods.

2 Related Work

2.1 Weakly Supervised RGB-T SOD

WSSOD remains at an early stage and has so far focused mainly on scribble supervision prompts. Liu et al. (2023) establish the first scribble-supervised RGB-T SOD setting on RGBT-S and propose a four-stage pipeline. Zhai et al. (2025) employ the Segment Anything Model (SAM) Kirillov et al. (2023) to optimize scribble-derived labels and progressively fuse cross-modal and cross-scale features for more reliable saliency learning. Liu et al. (2025) further build SSFam on SAM with modal-aware modulators and a siamese decoder to unify scribble-supervised learning across modality combinations. These methods mainly improve pseudo-label generation, but they are still restricted to scribble supervision prompts and typically treat all predicted regions and training samples as equally reliable, whereas ECL-Net supports point, box, and scribble supervision prompts and explicitly addresses region-level unreliability and sample-level difficulty.

2.2 Evidential Learning

The region- and sample-level unreliability discussed above requires a principled way to estimate which predictions are trustworthy during training. Evidential learning offers such a formulation by treating network outputs as collected evidence rather than point probabilities. Jsang (2018) formalize subjective logic as a framework for reasoning under uncertainty by relating belief, disbelief, and ignorance. Sensoy et al. (2018) instantiate this idea in deep learning for classification, where class-wise evidence is collected and predictive uncertainty is represented by a Dirichlet distribution. Tsiligkaridis (2021) further improve Dirichlet-based estimation with information-aware max-norm regularization under conflicting evidence. Amini et al. (2020) extend evidential modeling to regression through a Normal-Inverse-Gamma distribution that separates aleatoric and epistemic uncertainty for continuous outputs. Bao et al. (2021) apply evidential uncertainty to open-set action recognition for distinguishing in-distribution and out-of-distribution actions. Han et al. (2022) design trusted multi-view classification with dynamic evidential fusion across views. Li et al. (2024) generalize evidential learning to composite classification and quantify multiple uncertainty types through a hyper evidential formulation. Closer to dense prediction, Li et al. (2023b) develop region-based evidential deep learning for brain tumor segmentation, where local Dirichlet evidence yields voxel-wise segmentation uncertainty. Despite this progress, evidential uncertainty remains largely unexplored in WSSOD, where incomplete supervision prompts induce heterogeneous region reliability and uneven sample difficulty. ECL-Net bridges this gap by reusing a single Dirichlet uncertainty estimate for both region-level refinement in ECUR and sample-level curriculum scheduling in EUCM.

3 Methodology

We present ECL-Net, a unified multi-prompt WSSOD framework built on SAM3 (Carion et al., 2025) and equipped with ECUR and EUCM. In this framework, evidential uncertainty serves as a shared signal for both local feature refinement and global curriculum scheduling.

3.1 Problem Formulation and Overall Framework

Problem formulation Given RGB image $\mathbf{I}_{\mathrm{rgb}}$, thermal image $\mathbf{I}_{\mathrm{t}}$, and a weak supervision prompt $P_{\mathrm{w}} \in \{\text{point}, \text{box}, \text{scribble}\}$, we learn a saliency map $\hat{\mathbf{Y}} \in [0,1]^{H \times W}$ from partial spatial evidence through evidential refinement and curriculum scheduling.

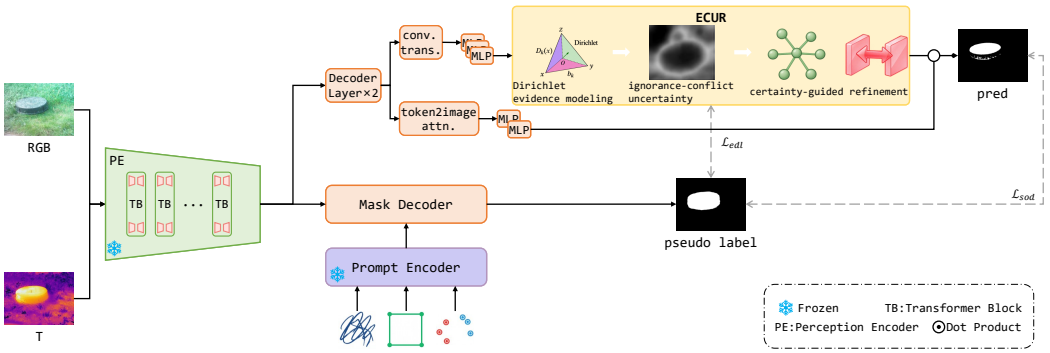


Figure 1: Overview of the proposed ECL-Net framework for WSSOD.

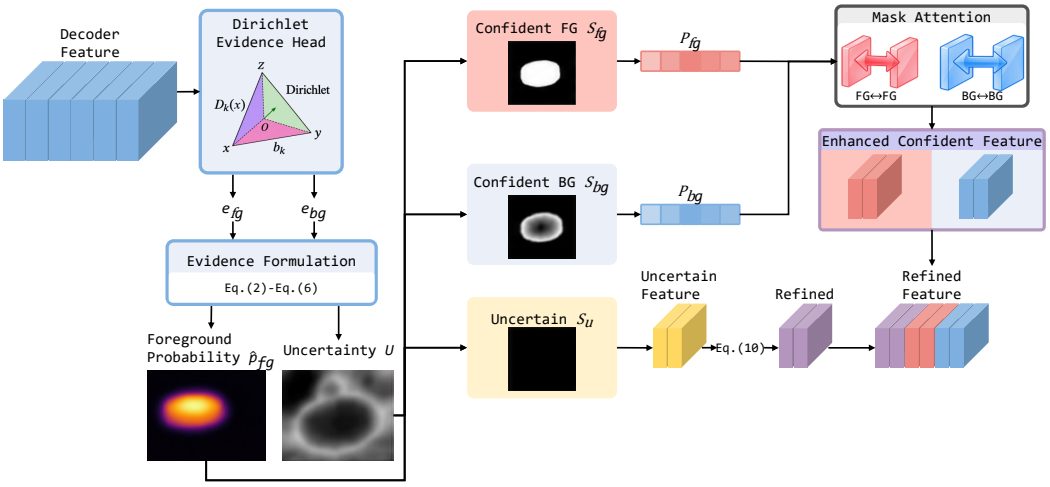


Figure 2: Architecture of ECUR.

Overall framework As shown in Figure 1, ECL-Net adopts a dual-branch SAM3 (Carion et al., 2025) design. A shared encoder (Hu et al., 2022) extracts and fuses modality-specific features $\{\mathbf{F}^{\ell}_{\mathrm{rgb}}\}_{\ell=1}^{L}$ and $\{\mathbf{F}^{\ell}_{\mathrm{t}}\}_{\ell=1}^{L}$ into $\{\mathbf{F}^{\ell}\}_{\ell=1}^{L}$, which feeds two decoders: $\mathrm{Dec_p}$ converts P_{w} into pseudo-labels $\hat{\mathbf{Y}}_{\mathrm{p}}$, and $\mathrm{Dec_{np}}$ with ECUR produces the saliency map $\hat{\mathbf{Y}}$. ECUR estimates pixel-wise epistemic uncertainty and refines ambiguous regions, while EUCM reuses the same uncertainty for sample-level easy-to-hard curriculum scheduling.

3.2 Evidential Certain-Uncertain Refinement Mechanism

Dirichlet evidence modeling. Given decoder feature $f(x)$ in Fig. 2, a Dirichlet Evidence Head predicts non-negative foreground/background evidence

$$e_{\mathrm{bg}}(x), e_{\mathrm{fg}}(x) = \mathrm{Softplus}\Big(\mathrm{Conv}_{1\times1}\big(\mathrm{ReLU}\big(\mathrm{GN}\big(\mathrm{Conv}_{3\times3}(f(x))\big)\big)\big)\Big), \tag{1}$$

which yield Dirichlet concentrations (Jsang, 2018) and total evidence (Sensoy et al., 2018)

$$\alpha_k(x) = e_k(x) + 1, k \in \{\mathrm{bg}, \mathrm{fg}\}, \qquad S(x) = \sum_{k\in\{\mathrm{bg},\mathrm{fg}\}} \alpha_k(x). \tag{2}$$

Ignorance–conflict uncertainty. For $K{=}2$ classes, epistemic uncertainty decomposes into ignorance (Zhou et al., 2023b)

$$I(x) = \frac{K}{S(x)}, \tag{3}$$

and conflict (Barhoumi et al., 2025)

$$b_k(x) = \frac{e_k(x)}{S(x)}, \qquad b_{\max}(x) = \max\big(b_{\mathrm{bg}}(x), b_{\mathrm{fg}}(x)\big), \tag{4}$$

$$C(x) = 1 - \frac{\big(b_{\max}(x) - b_{\mathrm{bg}}(x)\big) + \big(b_{\max}(x) - b_{\mathrm{fg}}(x)\big)}{K-1}. \tag{5}$$

Their probabilistic union (Zhou et al., 2022) yields

$$U(x) = I(x) + C(x) - I(x)C(x), \tag{6}$$

which separates confident and uncertain regions for refinement and EUCM scheduling.

Certainty-guided refinement. Confidence and Dirichlet mean (Tsiligkaridis, 2021) select reliable seeds

$$T(x) = 1 - \mathrm{Norm}(U(x)), \qquad \hat{p}_{\mathrm{fg}}(x) = \frac{\alpha_{\mathrm{fg}}(x)}{S(x)}, \tag{7}$$

$$\mathcal{S}_{\mathrm{fg}} = \{x \mid \hat{p}_{\mathrm{fg}}(x) > \tau_{\mathrm{fg}} \wedge T(x) > \tau_{\mathrm{c}}\}, \quad \mathcal{S}_{\mathrm{bg}} = \{x \mid \hat{p}_{\mathrm{fg}}(x) < \tau_{\mathrm{bg}} \wedge T(x) > \tau_{\mathrm{c}}\}. \tag{8}$$

In our main experiments, we set $(\tau_{\mathrm{fg}}, \tau_{\mathrm{bg}}, \tau_{\mathrm{c}})$ to $(0.8, 0.25, 0.75)$ for scribble, box, and point supervision. Appendix A reports the sensitivity analysis over alternative threshold choices. Certainty-weighted prototypes are updated by EMA with m=0.99 (Tarvainen & Valpola, 2017)

$$P_k = \frac{\sum_{x \in \mathcal{S}_k} T(x) f(x)}{\sum_{x \in \mathcal{S}_k} T(x) + \epsilon}. \tag{9}$$

For $x \in \mathcal{S}_{\mathrm{u}} = 1 - \mathcal{S}_{\mathrm{fg}} - \mathcal{S}_{\mathrm{bg}}$, features are pulled toward the more similar prototype,

$$f'(x) = (1 - \lambda_{\mathrm{u}}) f(x) + \lambda_{\mathrm{u}} P_{\hat{k}}, \quad \hat{k} = \arg\max_{k \in \{fg, bg\}} \mathrm{sim}_k(x), \quad \lambda_{\mathrm{u}} = \sigma(\eta), \tag{10}$$

where $\mathrm{sim}_k(x)$ denotes cosine similarity to P_k. On confident seeds $\mathcal{S}_{\mathrm{c}} = \mathcal{S}_{\mathrm{fg}} \cup \mathcal{S}_{\mathrm{bg}}$, masked self-attention (Cheng et al., 2022) with same-class mask $\boldsymbol{M}$ yields

$$f'(x) = f(x) + \lambda_{\mathrm{c}} f_{\mathrm{att}}(x), \quad f_{\mathrm{att}} = \mathrm{Unflatten}\left(\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{d}} + \Omega(\boldsymbol{M})\right)\boldsymbol{V}\right), \; x \in \mathcal{S}_{\mathrm{c}}, \tag{11}$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ are standard projections of flattened features $\boldsymbol{X} \in \mathbb{R}^{HW \times C}$, and $\Omega(\boldsymbol{M})$ masks disallowed token pairs with $-\infty$.

3.3 Evidential Uncertainty Curriculum Module

EUCM reuses the no-prompt uncertainty map $U(x)$ to schedule training from easy to hard samples. The difficulty score of sample i is

$$\rho(i) = \frac{1}{HW} \sum_{x} U(x), \tag{12}$$

where smaller $\rho(i)$ indicates lower difficulty. Training begins with a randomly selected warm-up pool. Whenever new samples are admitted, unselected samples are re-scored with the latest checkpoint, ranked by ascending $\rho(i)$, and selected until their cumulative difficulty reaches 20% of the total uncertainty across the training set. Curriculum updates are triggered when validation performance plateaus. This closes the loop from regional uncertainty estimation to sample scheduling, allowing stable RGB-T saliency priors to be learned before harder cases are introduced.

3.4 Loss Functions

Training uses three prompt-specific SOD losses and one shared evidential loss (Li et al., 2023b). The evidential term is

$$\mathcal{L}_{\mathrm{edl}} = \mathcal{L}_{\mathrm{dice}} + \lambda_{\mathrm{kl}} \mathcal{L}_{\mathrm{kl}}, \tag{13}$$

where $\mathcal{L}_{\mathrm{dice}}$ is the evidential Dice surrogate on Dirichlet parameters, $\mathcal{L}_{\mathrm{kl}}$ is the KL divergence to a non-informative Dirichlet prior, and $\lambda_{\mathrm{kl}} = 5 \times 10^{-4}$.

Table 1: Quantitative comparison with state-of-the-art methods under different supervision prompts. *Ours-S*, *Ours-B*, and *Ours-P* denote our model trained with scribble, box, and point supervision, respectively. The vs. $w/t/l$ column reports wins, ties, and losses against other methods on VT5000, measured by MAE. The best and second-best results across all methods are highlighted in red and cyan, respectively.

Method	Source	Modal	Supervision	VT821				VT1000				VT5000				vs. w/t/l
				$S\uparrow$	$F_w\uparrow$	$E\uparrow$	MAE$\downarrow$	$S\uparrow$	$F_w\uparrow$	$E\uparrow$	MAE$\downarrow$	$S\uparrow$	$F_w\uparrow$	$E\uparrow$	MAE$\downarrow$	
PRLNet(Zhou et al., 2023a)	TIP23	RGB-T	Full	0.916	0.869	0.938	0.026	0.893	0.822	0.920	0.037	0.918	0.886	0.956	0.023	11/0/3
SACNet(Wang et al., 2024)	TMM24	RGB-T	Full	0.908	0.861	0.939	0.025	0.940	0.928	0.976	0.014	0.915	0.891	0.959	0.021	12/1/1
SAMSOD(Liu et al., 2026)	TMM26	RGB-T	Full	0.924	0.897	0.954	0.021	0.945	0.938	0.979	0.012	0.929	0.914	0.967	0.017	14/0/0
WSSA(Zhang et al., 2020)	CVPR20	RGB-D	Scribble	0.821	0.703	0.845	0.050	0.863	0.797	0.894	0.042	0.782	0.656	0.831	0.062	2/0/12
SCWSSOD(Yu et al., 2021)	AAAI21	RGB	Scribble	0.840	0.761	0.864	0.044	0.870	0.828	0.903	0.039	0.808	0.725	0.850	0.052	3/0/11
DENet(Xu et al., 2022)	AAAI21	RGB-D	Scribble	0.703	0.568	0.722	0.071	0.724	0.626	0.748	0.081	0.641	0.469	0.649	0.082	1/0/13
MIRV(Li et al., 2023a)	TCSVT23	RGB	Scribble	0.516	0.214	0.412	0.112	0.499	0.208	0.399	0.142	0.475	0.122	0.339	0.114	0/0/14
SPB-Net(Li et al., 2023c)	TPAMI23	RGB-D	Scribble	0.860	0.726	0.872	0.048	0.913	0.837	0.932	0.035	0.849	0.716	0.875	0.048	4/0/10
RGBT-Scribble(Liu et al., 2023)	ICME23	RGB-T	Scribble	0.876	0.818	0.908	0.035	0.889	0.832	0.921	0.035	0.868	0.815	0.918	0.037	6/0/8
SSFam(Liu et al., 2025)	TMM25	RGB-T	Scribble	0.908	0.865	0.925	0.027	0.914	0.888	0.944	0.026	0.900	0.868	0.935	0.028	8/0/6
Ours-S	-	RGB-T	Scribble	0.926	0.893	0.952	0.024	0.946	0.934	0.977	0.015	0.927	0.902	0.962	0.021	12/1/1
NESS-Net(Zheng et al., 2026)	TCSVT26	RGB-T	Box	0.881	0.830	0.930	0.030	0.918	0.894	0.961	0.022	0.864	0.813	0.930	0.035	7/0/7
Ours-B	-	RGB-T	Box	0.909	0.869	0.944	0.030	0.936	0.918	0.972	0.017	0.911	0.882	0.953	0.025	9/1/4
PSOD(Gao et al., 2022)	AAAI22	RGB	Point	0.870	0.791	0.890	0.039	0.917	0.881	0.947	0.025	0.853	0.776	0.892	0.041	5/0/9
Ours-P	-	RGB-T	Point	0.920	0.897	0.952	0.024	0.938	0.931	0.972	0.018	0.912	0.891	0.951	0.025	9/1/4

For scribble supervision (Liu et al., 2025),

$$\mathcal{L}^{\mathrm{s}} = \mathcal{L}^{\mathrm{sod}} + \mathcal{L}_{\mathrm{edl}} = \mathcal{L}_{\mathrm{str}} + \mathcal{L}_{\mathrm{sal}} + \mathcal{L}_{\mathrm{sm}} + \mathcal{L}_{\mathrm{lsc}} + \mathcal{L}_{\mathrm{edl}}. \tag{14}$$

$\mathcal{L}_{\mathrm{str}}$ aligns no-prompt predictions with prompt-guided pseudo-labels; $\mathcal{L}_{\mathrm{sal}}$ applies sparse scribble supervision; $\mathcal{L}_{\mathrm{sm}}$ and $\mathcal{L}_{\mathrm{lsc}}$ enforce edge-aware smoothness and local saliency coherence.

For box supervision (Tian et al., 2021),

$$\mathcal{L}^{\mathrm{b}} = \mathcal{L}^{\mathrm{sod}} + \mathcal{L}_{\mathrm{edl}} = \mathcal{L}_{\mathrm{proj}} + \mathcal{L}_{\mathrm{pair}} + \mathcal{L}_{\mathrm{edl}}. \tag{15}$$

$\mathcal{L}_{\mathrm{proj}}$ matches horizontal and vertical saliency projections to the box; $\mathcal{L}_{\mathrm{pair}}$ encourages local consistency via RGB feature affinity within the box.

For point supervision,

$$\mathcal{L}^{\mathrm{p}} = \mathcal{L}^{\mathrm{sod}} + \mathcal{L}_{\mathrm{edl}} = \mathcal{L}_{\mathrm{str}} + \mathcal{L}_{\mathrm{sal}} + \mathcal{L}_{\mathrm{lsc}} + \mathcal{L}_{\mathrm{edl}}. \tag{16}$$

4 Experiments

4.1 Experimental Setup

We train on RGBT-S (Liu et al., 2023) and evaluate on VT821 (Wang et al., 2018), VT1000 (Tu et al., 2019), and VT5000 (Tu et al., 2022). Scribble labels follow the dataset; box prompts are tight ground-truth rectangles; point prompts are sampled from foreground scribbles. We report S (Fan et al., 2017), F_w (Margolin et al., 2014), E (Fan et al., 2018), and MAE (Perazzi et al., 2012).

All images are resized to 1008×1008. We fine-tune SAM3 (Carion et al., 2025) with Adam (Kingma & Ba, 2014) for 30 epochs (learning rate 5×10^{-5}, batch size 1) on one NVIDIA RTX 3090 GPU ($\sim$16 hours).

4.2 Comparison with State-of-the-Art Methods

Quantitative evaluation Table 1 compares fully supervised and weakly supervised methods. Among weakly supervised approaches, Ours-S gives the best results on all three test sets and is ahead of recent WSSOD methods such as SSFam and RGBT-Scribble. The same trend appears with coarser supervision prompts, where Ours-B is better than NESS-Net on every metric. To our knowledge, no point-supervised RGB-T SOD method has been reported so far, so we compare Ours-P against PSOD, a point-supervised RGB SOD method, and Ours-P achieves better results on all metrics. This suggests ECUR and EUCM remain useful when little spatial information is given. Ours-S is also close to fully supervised performance. On VT5000 it reaches $S = 0.927$, which is within 0.002 of the strongest fully supervised method SAMSOD (0.929), with $\mathrm{MAE} = 0.021$ on par with SACNet. Overall the gap between weak and full supervision is substantially reduced.

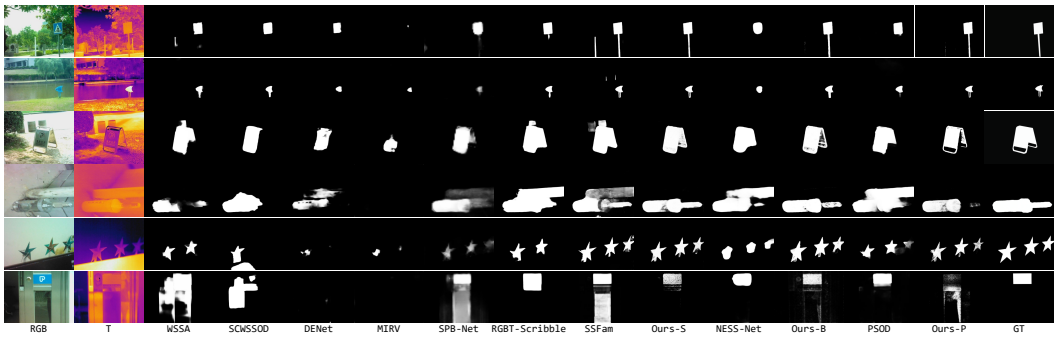


Figure 3: Qualitative comparison with state-of-the-art methods under point, box, and scribble supervision prompts.

Table 2: Ablation study of ECUR and EUCM under point, box, and scribble supervision prompts. Checkmarks indicate whether the SAM3 baseline (B), ECUR, and EUCM are enabled. Best results within each supervision group are highlighted in bold.

Supervision	Modules			VT821				VT1000				VT5000			
	B	ECUR	EUCM	$S\uparrow$	$F_w\uparrow$	$E\uparrow$	MAE$\downarrow$	$S\uparrow$	$F_w\uparrow$	$E\uparrow$	MAE$\downarrow$	$S\uparrow$	$F_w\uparrow$	$E\uparrow$	MAE$\downarrow$
Scribble	✓			0.918	0.873	0.938	0.028	0.943	0.928	0.973	0.016	0.910	0.871	0.943	0.028
	✓	✓		**0.927**	0.891	0.948	0.025	0.944	0.931	0.974	0.016	0.918	0.888	0.952	0.025
	✓		✓	0.924	0.889	0.947	**0.023**	0.944	0.931	0.975	**0.015**	0.925	0.897	0.957	**0.021**
	✓	✓	✓	0.926	**0.893**	**0.952**	0.024	**0.946**	**0.934**	**0.977**	**0.015**	**0.927**	**0.902**	**0.962**	**0.021**
Box	✓			0.772	0.614	0.798	0.095	0.815	0.696	0.843	0.079	0.749	0.573	0.762	0.107
	✓	✓		0.811	0.677	0.843	0.070	0.848	0.748	0.879	0.058	0.781	0.623	0.804	0.084
	✓		✓	0.906	0.858	0.937	**0.028**	0.931	0.906	0.966	0.018	0.902	0.855	0.939	0.029
	✓	✓	✓	**0.909**	**0.869**	**0.944**	0.030	**0.936**	**0.918**	**0.972**	**0.017**	**0.911**	**0.882**	**0.953**	**0.025**
Point	✓			0.500	0.264	0.425	0.116	0.492	0.288	0.434	0.145	0.504	0.270	0.432	0.108
	✓	✓		0.599	0.449	0.588	0.103	0.556	0.408	0.548	0.135	0.589	0.441	0.583	0.098
	✓		✓	0.910	0.876	0.937	0.035	0.931	0.919	0.964	0.022	0.901	0.872	0.939	0.031
	✓	✓	✓	**0.920**	**0.897**	**0.952**	**0.024**	**0.938**	**0.931**	**0.972**	**0.018**	**0.912**	**0.891**	**0.951**	**0.025**

Qualitative evaluation Figure 3 shows examples that contain difficult factors such as Big Salient Object, Center Bias, Image Clutter, Multiple Salient Objects, Small Salient Object, Thermal Crossover, RGB Modality Challenge, and Thermal Modality Challenge. While other weakly supervised methods often break the foreground into parts, activate background clutter, or blur the boundary, ECL-Net keeps the object intact and the background quiet. The observed behavior is consistent with the design of ECL-Net. ECUR identifies uncertain regions from Dirichlet evidence and corrects them using foreground and background prototypes, while EUCM stabilizes training by ordering samples according to sample-level difficulty scores derived from ignorance and conflict uncertainties. Together, they support more reliable cross-modal saliency learning under sparse supervision prompts, yielding complete foreground and suppressed background responses.

4.3 Ablation Studies

Overall ablation Table 2 reports results for scribble, box, and point supervision prompts when each proposed component is added to the SAM3 baseline B. Adding either ECUR or EUCM improves accuracy on all three benchmarks. With scribble prompts, which already carry richer shape cues, B+ECUR+EUCM gives the best overall balance. The effect is larger under coarser supervision prompts. With box prompts, where the rectangle inevitably contains background, the baseline is low and EUCM brings most of the gain by keeping hard examples from dominating early training. With point prompts, where supervision is a single location, the baseline is weakest and both modules help, with EUCM providing the main stabilization and ECUR adding a further correction.

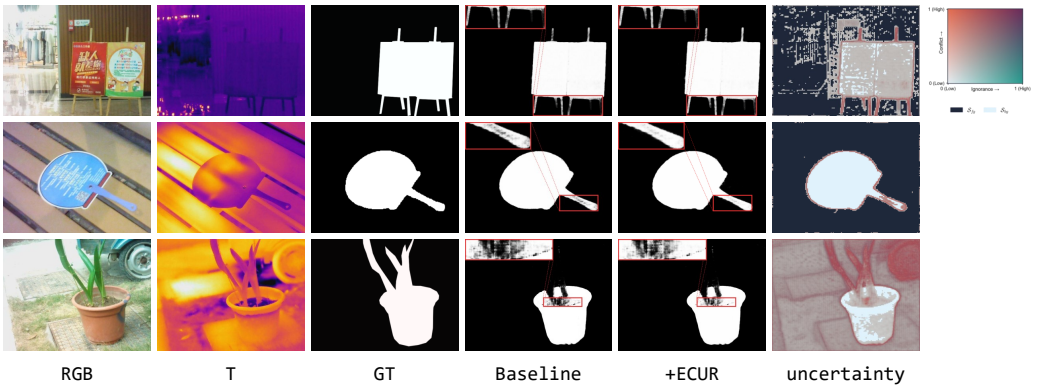


Figure 4: Qualitative ablation of ECUR under point, box, and scribble supervision prompts. From top to bottom, the three rows correspond to scribble, box, and point supervision, respectively.

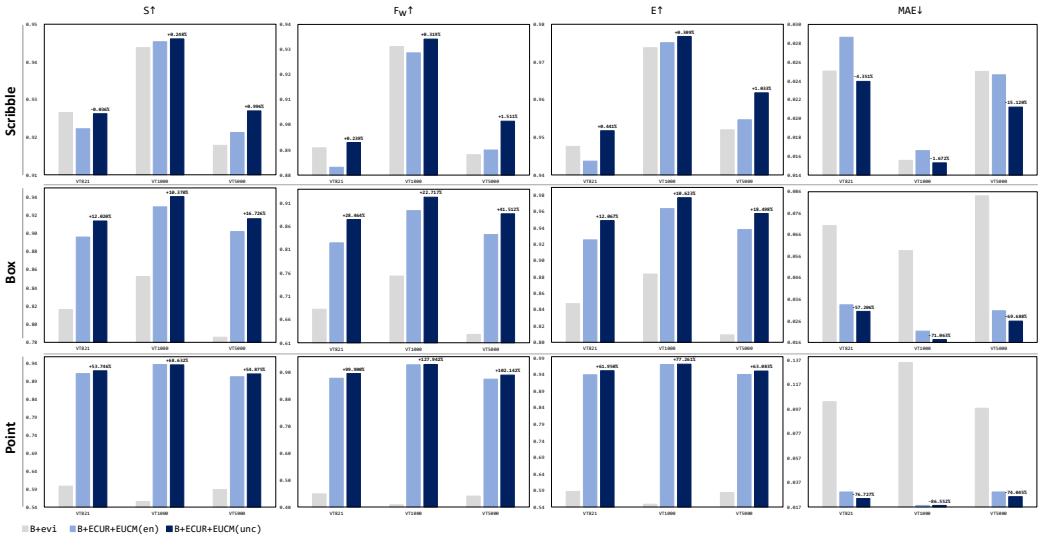


Figure 5: Quantitative ablation of EUCM under point, box, and scribble supervision prompts. en ranks sample difficulty by the entropy of the predicted foreground probability, while unc ranks samples using the evidential uncertainty in Eq. (12).

Together they address local feature noise and global training order, which explains the consistent improvement across supervision prompts.

Ablation on ECUR Figure 4 compares predictions with and without ECUR under point, box, and scribble supervision prompts. ECUR brings the mask closer to the ground truth in all cases, but the correction is not uniform. Foreground is fixed more easily, while background confidence builds more slowly, especially with scribble and point labels. Weak supervision prompts cover only parts of the foreground, so the model sees little direct evidence for background and its epistemic uncertainty there falls slowly. Prototypes are also updated by certainty-weighted averaging with EMA, so each step is deliberately small. This avoids instability from noisy updates but lets the background prototype adapt gradually. With point supervision a single location often makes RGB and thermal features disagree locally, so the conflict part of uncertainty dominates, while scribble and box cases are dominated by ignorance.

Ablation on EUCM Figure 5 tests EUCM using the evidential uncertainty schedule (unc) against an entropy baseline (en) that ranks sample difficulty by the entropy of the predicted foreground

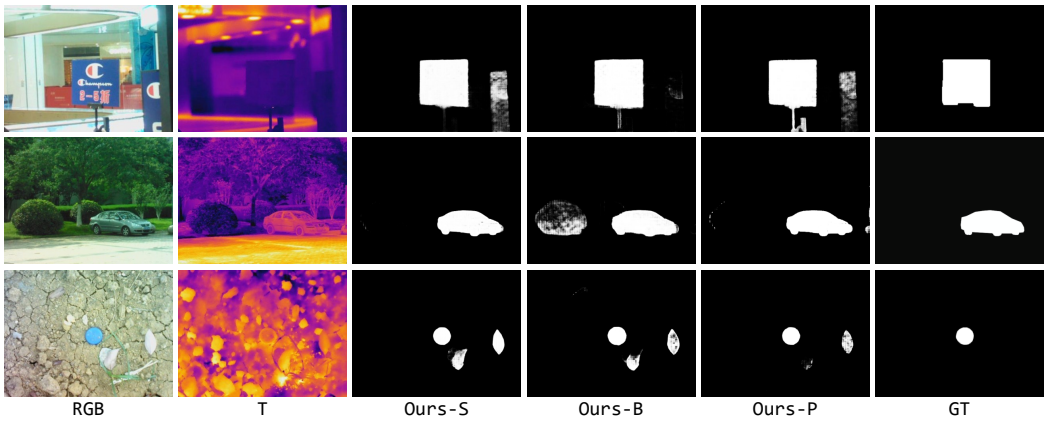


Figure 6: Failure cases of ECL-Net.

probability. The unc schedule is better on all three supervision prompts. Entropy reflects only how spread the probability is. EUCM instead aggregates pixel-level ignorance and conflict uncertainties from Dirichlet evidence into sample-level difficulty scores, so the ranking follows what the model actually does not know.

4.4 Failure Cases and Limitations

Figure 6 shows representative failure cases of ECL-Net, where a faint residual remains near object boundaries or in background regions with a foreground-like appearance. These artifacts mainly arise from the limited spatial evidence inherent to weak supervision prompts. This suggests that boundary fidelity remains the main bottleneck of ECL-Net.

5 Conclusion

We proposed ECL-Net, a unified multi-prompt WSSOD framework on SAM3 that couples region-level evidential refinement with sample-level curriculum scheduling. ECUR models Dirichlet evidence to separate ignorance and conflict uncertainties and refine unreliable regions, while EUCM reuses the same estimate to rank samples and admit them in an easy-to-hard order. Experiments on VT821, VT1000, and VT5000 show consistent gains over existing WSSOD methods and performance close to fully supervised counterparts. Weak supervision prompts still leave faint boundary responses in complex scenes; future work may exploit complementary RGB-T saliency priors to compensate for sparse supervision.

AI use statement

In this work, we used generative AI tools for polishing manuscript text and assisting with LaTeX formatting. We did not use generative AI tools for developing the core research ideas, designing the method, running experiments, or generating experimental results. All AI-assisted text was reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility statement

We describe the datasets, evaluation metrics, and implementation details in Section 4.1, including the RGBT-S training set and the VT821, VT1000, and VT5000 benchmarks, and the optimizer, learning rate, batch size, and EUCM curriculum schedule. Hyperparameter sensitivity for ECUR seed selection and scribble supervision prompt sampling analysis are provided in Appendices A and B. Anonymized source code and trained model weights will be released as supplementary material to support reproduction of the reported results.

ETHICS STATEMENT

This work uses only publicly available RGB-T saliency datasets and does not involve human subjects, sensitive personal data, or deceptive practices. The proposed method targets generic salient object detection, and we are not aware of foreseeable direct harmful applications beyond the general dual-use considerations shared by standard visual perception research.

REFERENCES

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937, 2020.

Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13349–13358, 2021.

Samia Barhoumi, Imene Khanfir Kallel, Eloi Bosse, and Basel Solaiman. An empirical survey-type analysis of uncertainty measures for the fusion of crisp and fuzzy bodies of evidence. *Information Fusion*, 121:103106, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pp. 4548–4557, 2017.

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018.

Shuyong Gao, Wei Zhang, Yan Wang, Qianyu Guo, Chenglong Zhang, Yangji He, and Wenqiang Zhang. Weakly-supervised salient object detection using point supervision. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 670–678, 2022.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2551–2566, 2022.

Jia Hou, Hongfa Wen, Qiang Zhao, Junbo Ma, Zongpeng Li, Shuai Wang, and Chenggang Yan. A comprehensive survey on the research and development of rgb-t salient object detection. *Computational Visual Media*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Audun Jsang. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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 *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Aixuan Li, Yuxin Mao, Jing Zhang, and Yuchao Dai. Mutual information regularization for weakly-supervised rgb-d salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):397–410, 2023a.

Changbin Li, Kangshuo Li, Yuzhe Ou, Lance Kaplan, Audun Jøsang, Jin-Hee Cho, Dong Hyun Jeong, and Feng Chen. Hyper evidential deep learning to quantify composite classification uncertainty. In *International Conference on Learning Representations*, volume 2024, pp. 8826–8860, 2024.

Hao Li, Yang Nan, Javier Del Ser, and Guang Yang. Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation. *Neural Computing and Applications*, 35(30):22071–22085, 2023b.

Long Li, Junwei Han, Nian Liu, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Robust perception and precise segmentation for scribble-supervised rgb-d saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1): 479–496, 2023c.

Yi Li, Zhanghui Kuang, Liyang Liu, Yimin Chen, and Wayne Zhang. Pseudo-mask matters in weakly-supervised semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6944–6953. IEEE, 2021.

Zhengyi Liu, Xiaoshen Huang, Guanghui Zhang, Xianyong Fang, Linbo Wang, and Bin Tang. Scribble-supervised rgb-t salient object detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 2369–2374. IEEE, 2023.

Zhengyi Liu, Sheng Deng, Xinrui Wang, Linbo Wang, Xianyong Fang, and Bin Tang. Ssfam: Scribble supervised salient object detection family. *IEEE Transactions on Multimedia*, 2025.

Zhengyi Liu, Xinrui Wang, Xianyong Fang, Zhengzheng Tu, and Linbo Wang. Samsod: Rethinking sam optimization for rgb-t salient object detection. *IEEE Transactions on Multimedia*, 2026.

Huayu Mai, Rui Sun, Yujia Chen, Wangkai Li, Bingzhou Wang, Aibing Li, Zhangyu He, and Yuan Wang. From softmax to dirichlet: Evidential learning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27697–27707, 2026.

Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2014.

Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 733–740. IEEE, 2012.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5439–5448. IEEE, 2021.

Theodoros Tsiligkaridis. Information aware max-norm dirichlet networks for predictive uncertainty estimation. *Neural Networks*, 135:105–114, 2021.

Zhengzheng Tu, Tian Xia, Chenglong Li, Xiaoxiao Wang, Yan Ma, and Jin Tang. Rgb-t image saliency detection via collaborative graph learning. *IEEE Transactions on Multimedia*, 22(1): 160–173, 2019.

Zhengzheng Tu, Yan Ma, Zhun Li, Chenglong Li, Jieming Xu, and Yongtao Liu. Rgbt salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia*, 25: 4163–4176, 2022.

Guizhao Wang, Chenglong Li, Yunpeng Ma, Aihua Zheng, Jin Tang, and Bin Luo. Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In *Chinese Conference on Image and Graphics Technologies*, pp. 359–369. Springer, 2018.

Kunpeng Wang, Danying Lin, Chenglong Li, Zhengzheng Tu, and Bin Luo. Alignment-free rgbt salient object detection: Semantics-guided asymmetric correlation network and a unified benchmark. *IEEE Transactions on Multimedia*, 26:10692–10707, 2024.

Mianzhao Wang, Fan Shi, Xu Cheng, Chen Jia, and Shengyong Chen. Uncertainty-aware modality fusion for unaligned rgb-t salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41489–41498, 2026.

Binwei Xu, Qiuping Jiang, Xing Zhao, Chenyang Lu, Haoran Liang, and Ronghua Liang. Multi-dimensional exploration of segment anything model for weakly supervised video salient object detection. *IEEE Transactions on circuits and systems for video technology*, 35(4):2987–2998, 2024.

Yunqiu Xu, Xin Yu, Jing Zhang, Linchao Zhu, and Dadong Wang. Weakly supervised rgb-d salient object detection with prediction consistency training and active scribble boosting. *IEEE Transactions on Image Processing*, 31:2148–2161, 2022.

Siyue Yu, Bingfeng Zhang, Jimin Xiao, and Eng Gee Lim. Structure-consistent weakly supervised salient object detection with local saliency coherence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 3234–3242, 2021.

Sulan Zhai, Chengzhuang Liu, Zhengzheng Tu, Chenglong Li, and Liuxuanqi Gao. Weakly supervised rgbt salient object detection via sam-guided label optimization and progressive cross-modal cross-scale fusion. *Information Fusion*, 120:103048, 2025.

Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12546–12555, 2020.

Yunping Zheng, Hao Li, ShiQiang Shu, Ke Xu, and Mudar Sarem. Ness-net: An namlab edge-guided and scribble-supervised swin-transformer net for rgb-d salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Heng Zhou, Chunna Tian, Zhenxi Zhang, Chengyang Li, Yuxuan Ding, Yongqiang Xie, and Zhongbo Li. Position-aware relation learning for rgb-thermal salient object detection. *IEEE Transactions on Image Processing*, 32:2593–2607, 2023a.

Qianli Zhou, Éloi Bossé, and Yong Deng. Modeling belief propensity degree: Measures of evenness and diversity of belief functions. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(5):2851–2862, 2022.

Qianli Zhou, Witold Pedrycz, Yingying Liang, and Yong Deng. Information granule based uncertainty measure of fuzzy evidential distribution. *IEEE Transactions on Fuzzy Systems*, 31(12): 4385–4396, 2023b.

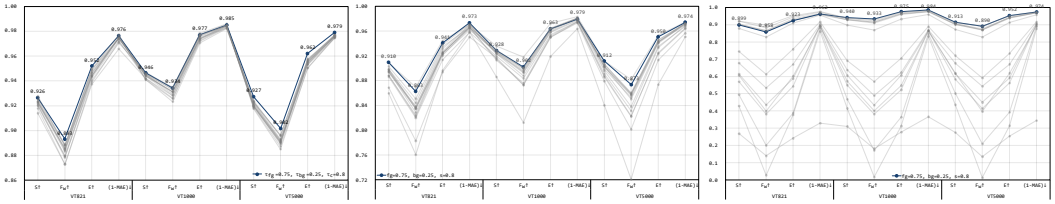


Figure 7: Hyperparameter sensitivity under scribble, box, and point supervision. Each curve corresponds to one configuration of $(\tau_{\text{fg}}, \tau_{\text{bg}}, \tau_{\text{c}})$. The best-performing configuration is highlighted in blue, while the remaining configurations are shown in gray.

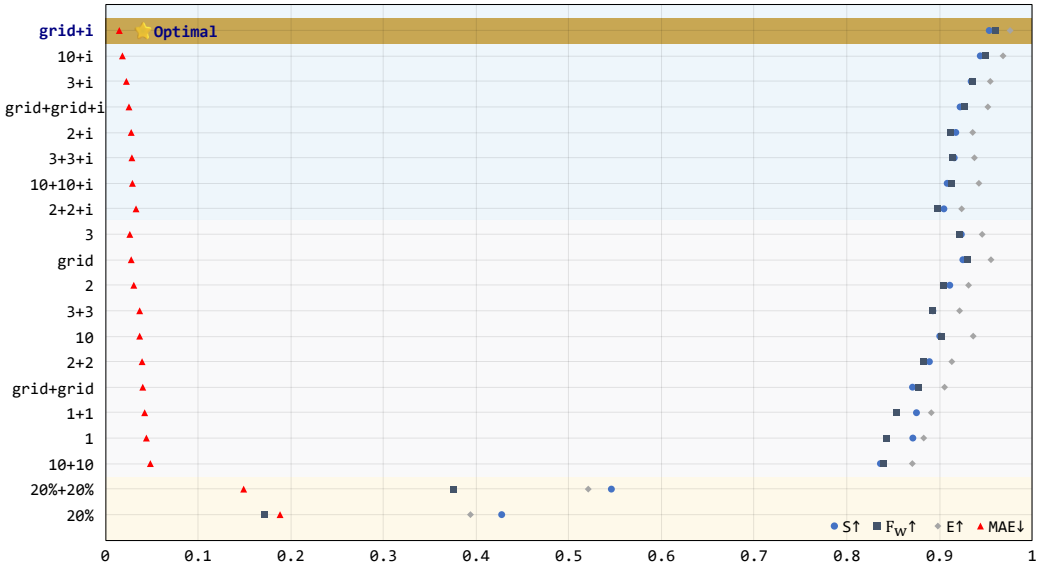


Figure 8: Pseudo-label quality under different scribble sampling strategies. The sampling strategies are organized into three categories: *iterative sampling*, *fixed-number sampling*, and *proportional sampling*. In the sampling notation, *+i* denotes iterative sampling, *a+b* denotes joint foreground and background sampling, while *a* denotes foreground-only sampling. For each component *a* or *b*, the sampling strategies can be grid-based 9-cell sampling (*grid*), fixed-point sampling with x points (x), or proportional sampling with $x\%$ points ($x\%$).

This appendix provides supplementary experimental details, including hyperparameter sensitivity analysis for ECUR seed selection and scribble sampling analysis for supervision prompt generation.

A Hyperparameter Sensitivity of ECUR Seed Selection

We analyze the foreground probability threshold τ_{fg}, background probability threshold τ_{bg}, and confidence threshold τ_{c} used for ECUR seed selection in Eq. (8). Because ECL-Net is trained as a *unified model* shared across point, box, and scribble supervision prompts, we search for one threshold configuration that remains competitive under all three prompt types instead of tuning supervision-specific settings. In all sweeps, we fix $\tau_{\text{fg}} - \tau_{\text{bg}} = 0.5$ and vary $\tau_{\text{fg}}/\tau_{\text{bg}}$ over $\{0.7/0.2,\ 0.75/0.25,\ 0.8/0.3\}$ and τ_{c} over $\{0.5,\ 0.6,\ 0.7,\ 0.8,\ 0.9\}$. Figure 7 reports the sensitivity results under the three supervision types. We adopt $(\tau_{\text{fg}}, \tau_{\text{bg}}, \tau_{\text{c}}) = (0.8, 0.25, 0.75)$ in all experiments.

B SCRIBBLE SAMPLING ANALYSIS

Figure 8 evaluates pseudo-label quality on VT5000-Train for different ways of sampling scribble-derived supervision prompts. Prompts that cover the foreground more evenly tend to produce cleaner pseudo-labels, and the foreground 9-grid iterative sampling (Xu et al., 2024) is best on the full metric set. It divides the foreground scribble region into a 3×3 grid and draws one point from each cell that contains foreground randomly, then combines those points with the mask predicted in the previous round as prompts for the next iteration. Better pseudo-labels in this stage give the no-prompt decoder $\mathrm{Dec}_{\mathrm{np}}$ a more reliable target, which lowers optimization noise and stabilizes training.