NON-EQUILIBRIUM DIFFUSION PSEUDO-LABELLING AND RELATIVE PREFERENCE OPTIMISATION FOR UNSUPERVISED CAMOUFLAGED OBJECT DETECTION

Anonymous authors
Paper under double-blind review

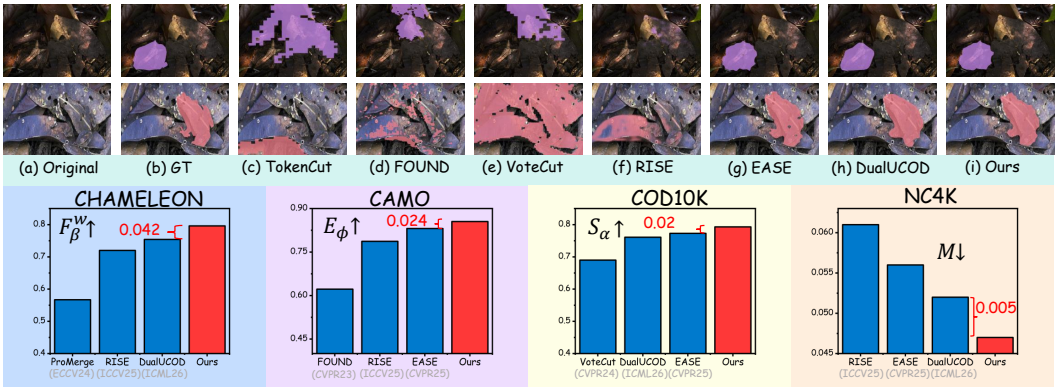
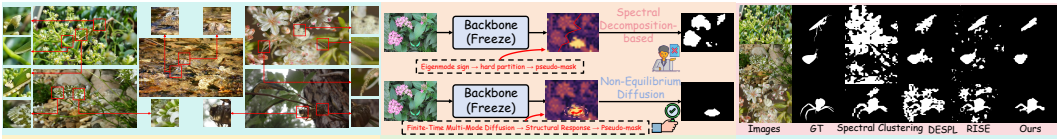


Figure 1: Qualitative and quantitative comparisons with different methods. Visual comparisons are conducted with TokenCut (Wang et al., 2022b), FOUND (Siméoni et al., 2023), VoteCut (Arica et al., 2024), RISE (Du et al., 2025b), EASE (Du et al., 2025a), and DualUCOD (Liu et al., 2026a). Our method yields more complete camouflaged objects with cleaner boundaries and reduced background interference in challenging scenes. Quantitative performance comparisons are further reported on CHAMELEON, CAMO, COD10K, and NC4K using F_β^w, E_ϕ, S_α, and M, respectively.

ABSTRACT

Unsupervised Camouflaged Object Detection (UCOD) is challenged by extreme visual entanglement between objects and surroundings. Existing pseudo-labelling methods rely on the Fiedler vector or a few low-frequency directions for binary partitioning, whereas *intra-object multi-camouflage* disperses a coherent foreground across weakly coupled structures, causing incomplete pseudo-labels whose noise further biases representation learning. To address these problems, we propose **DES-UCOD**, a novel **D**iscover-**E**xplore-**S**pecialise framework for **U**nsupervised **C**amouflaged **O**bject **D**etection that shifts from passive label fitting to active object discovery. Firstly, to resolve feature interference under multi-camouflage conditions and provide complete pseudo-supervision, we propose the Non-Equilibrium Diffusion Pseudo-labelling (NEDP) strategy, which captures transient structural dynamics prior to global equilibrium to theoretically unify fragmented camouflage cues and *discover* robust structural anchors to generate high-quality pseudo-labels. Secondly, to mitigate the negative impact of pseudo-label noise and significantly alleviate confirmation bias, we propose the Stochastic Relative Preference Optimisation (S-RPO) strategy, an active reward-driven sampling mechanism that compels the model to *explore* decision boundaries that are highly robust to diverse visual variations during training. Finally, to enhance the model's capacity for handling diverse camouflage characteristics, we propose a Spatially Conditioned Mixture of Experts (SC-MoE) decoder, which *specialises* in routing spatially heterogeneous features at each location, ensuring that distinct experts fulfil their respective roles to generate accurate predictions. Extensive experiments demonstrate that our model consistently outperforms UCOD methods across multiple datasets to achieve state-of-the-art performance.

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.



(a) Visualization of camouflaged objects under various camouflage conditions.

(b) Comparison between our method and previous pseudo-label-based methods.

(c) Visual comparison of pseudo-labels between our method and other methods.

Figure 2: **Comparison of pseudo-label generation for camouflaged object detection.** (a) Camouflaged objects under challenging conditions. (b) Existing spectral methods use limited low-frequency modes for partitioning, while our non-equilibrium diffusion integrates multiple modes to recover more complete foreground structures. (c) Pseudo-label comparisons showing improved completeness and cleaner boundaries.

1 INTRODUCTION

Camouflaged Object Detection (COD) (Fan et al., 2020a; He et al., 2023a) aims to identify targets visually indistinguishable from their surroundings. Given the extreme foreground-background similarities in colour, texture, and morphology, this task remains a formidable challenge. Although COD exhibits immense potential across military (Lee et al., 2024), medical (Fan et al., 2020b), and industrial (Fang et al., 2020) domains, fully supervised methods incur prohibitive pixel-level annotation costs. To circumvent these exorbitant expenses, researchers are increasingly shifting toward unsupervised learning paradigms to enable scalable real-world applications.

Unsupervised Camouflaged Object Detection (UCOD) has made steady progress in annotation-free object discovery. Existing methods can be broadly divided into self-learning Optimisation and prototype retrieval, as compared qualitatively and quantitatively in Fig. 1. Self-learning methods (Zhang & Wu, 2023; Shou et al., 2025; Liu et al., 2026b; Yan et al., 2025; Liu et al., 2026a) iteratively refine predictions using self-generated supervision. For example, UCOD-DPL (Yan et al., 2025) combines heuristic pseudo-labels with teacher predictions, while DualUCOD (Liu et al., 2026a) adopts dual-eigenvector pseudo-labelling with boundary and contrastive refinement. However, existing spectral approaches mainly rely on the Fiedler vector or a few low-frequency directions, which are insufficient for *intra-object multi-camouflage*, where different object parts blend into distinct surroundings and form weakly coupled structures in feature space. As shown in Fig. 2, limited spectral partitions may therefore recover only partial foreground regions, producing incomplete and noisy pseudo-labels. Prototype-retrieval methods instead introduce dataset-level references, such as environmental prototypes in EASE (Du et al., 2025a) and clustering-based multi-view retrieval in RISE (Du et al., 2025b), but become ambiguous when foreground and background share similar prototype affinities. Thus, self-learning is constrained by incomplete initial structures, whereas prototype retrieval depends on uncertain cross-instance matches. Motivated by *scale heterogeneity* (Nadler & Galun, 2006) and *finite-step smoothing* (Keriven, 2022), we exploit transient graph dynamics to aggregate multiple low-frequency responses before global equilibrium, thereby discovering complementary structures within each image without external prototype matching. We further introduce reward-guided relative exploration to prevent residual pseudo-label errors from being repeatedly reinforced and to model spatially heterogeneous camouflage patterns. Accordingly, we propose the *Discover-Explore-Specialise* framework to jointly address incomplete structural evidence, supervisory bias, and ambiguous cross-instance matching. (*See Appendix A.1 for a broader discussion of related work*).

In this work, we propose *DES-UCOD*, a *Discover-Explore-Specialise* framework for robust unsupervised camouflaged object detection. Firstly, Non-Equilibrium Diffusion pseudo-labelling (NEDP) captures transient graph responses before equilibrium to *discover* complementary structures and generate reliable pseudo-labels under *intra-object multi-camouflage*. Secondly, Stochastic Relative Preference Optimisation (S-RPO) uses reward-guided sampling and relative feedback to *explore* robust segmentation decisions while reducing confirmation bias from noisy supervision. Finally, the Spatially Conditioned Mixture of Experts (SC-MoE) decoder adaptively routes heterogeneous spatial features, enabling the model to *specialise* in diverse camouflage patterns. Extensive experiments on four benchmarks demonstrate the effectiveness of DES-UCOD.

Our main contributions are summarized as follows:

- We propose the *DES-UCOD* framework for UCOD featuring the NEDP strategy that leverages non-equilibrium diffusion to capture transient graph responses before global equilib-

rium and aggregate complementary low-frequency structures beyond limited spectral partitions for generating more complete pseudo-labels under intra-object multi-camouflage.

- To learn from the resulting richer yet imperfect supervision, we develop the S-RPO, which combines reward-guided stochastic sampling with relative preference learning to mitigate overfitting to residual labeling errors and reduce confirmation bias.
- We design the SC-MoE decoder that adaptively routes local features among experts according to spatial cues, enabling Specialised processing of heterogeneous camouflage patterns and improving foreground-background discrimination.
- Extensive experiments on four publicly available benchmarks demonstrate the effectiveness of DES-UCOD and its superior performance over existing UCOD methods.

2 Spectral Motivation for Multi-Mode Structural Discovery

This section provides an oracle spectral motivation for NEDP's multi-mode construction. The unknown foreground indicator is used only to characterize foreground-associated spectral energy and is never accessed by the practical algorithm. The analysis motivates finite-time spectral weighting without claiming that the label-free node response directly recovers the foreground mask. Its quality is evaluated in Appendix A.2.3.

2.1 Graph Formulation and Spectral Dispersion

Given an unlabelled image, we construct an undirected affinity graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ over N patches. We assume a connected graph with positive node degrees, as ensured by the affinity construction in Section 3.1. Let $\mathbf{D}$ be the degree matrix and $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ the symmetric normalised Laplacian. With distinct eigenvalues $0 = \mu_1 < \mu_2 < \cdots < \mu_Q$ and eigenspace projectors $\mathbf{P}_r$,

$$\mathbf{L} = \sum_{r=1}^{Q} \mu_r \mathbf{P}_r, \qquad \sum_{r=1}^{Q} \mathbf{P}_r = \mathbf{I}. \tag{1}$$

For oracle analysis, let $\mathbf{y} \in \{0,1\}^N$ denote the unknown foreground indicator. Define $\mathbf{g} = \mathbf{D}^{\frac{1}{2}}\mathbf{y}$ and its non-stationary component $\bar{\mathbf{g}} = (\mathbf{I} - \mathbf{P}_1)\mathbf{g}$. Neither is used for pseudo-label generation.

Definition 1 (Multi-Camouflage Spectral Dispersion)**.** *Assume $\bar{\mathbf{g}} \neq \mathbf{0}$ and $\|\mathbf{\Pi}_M \bar{\mathbf{g}}\|_2 > 0$. Let $\mathbf{\Pi}_M = \sum_{r=2}^{M+1} \mathbf{P}_r$ denote the retained low-frequency subspace. Its structural coverage and Fiedler-relative dispersion are*

$$C_M = \frac{\|\mathbf{\Pi}_M \bar{\mathbf{g}}\|_2^2}{\|\bar{\mathbf{g}}\|_2^2}, \qquad \mathcal{D}_M = 1 - \frac{\|\mathbf{P}_2 \bar{\mathbf{g}}\|_2^2}{\|\mathbf{\Pi}_M \bar{\mathbf{g}}\|_2^2}. \tag{2}$$

Here, C_M measures the foreground-associated energy retained in the low-frequency subspace, while $\mathcal{D}_M$ measures the fraction beyond the Fiedler direction. Thus, high C_M and non-negligible $\mathcal{D}_M$ characterize an oracle multi-mode regime. Both are analytical diagnostics only.

Proposition 1 (Subspace Stability of Weakly Coupled Camouflage Parts)**.** *Consider an ideal graph with q connected components, consisting of one background component and $q-1$ internally connected camouflage parts, whose normalised Laplacian is $\mathbf{L}_0 = \mathrm{diag}(\mathbf{L}_1, \ldots, \mathbf{L}_q)$. Let $\gamma = \lambda_{q+1}(\mathbf{L}_0) = \min_{1 \leq r \leq q} \lambda_2(\mathbf{L}_r) > 0$ and $\mathbf{L} = \mathbf{L}_0 + \mathbf{E}$ with $\|\mathbf{E}\|_2 \leq \eta < \gamma/2$. If $\mathcal{U}_0$ is the nullspace of $\mathbf{L}_0$, $\mathcal{U}_q$ is spanned by the q smallest eigenvectors of $\mathbf{L}$, and $\mathbf{\Pi}_q$ is the orthogonal projector onto $\mathcal{U}_q$, then the Davis–Kahan theorem (Davis & Kahan, 1970) gives*

$$\|\sin\Theta(\mathcal{U}_q, \mathcal{U}_0)\|_2 \leq \frac{\eta}{\gamma - \eta} \leq \frac{2\eta}{\gamma}, \qquad \frac{\|(\mathbf{I} - \mathbf{\Pi}_q)\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \frac{2\eta}{\gamma}, \quad \textit{for any } \mathbf{x} \in \mathcal{U}_0. \tag{3}$$

Proposition 1 shows that weakly coupled components remain represented in a stable low-dimensional spectral subspace. It does not identify foreground components or guarantee recovery from any single eigenvector.

2.2 Finite-Time Structural Response and Oracle Reconstruction

We aggregate complementary non-stationary eigenspaces through finite-time graph diffusion to capture richer low-frequency structural information before global equilibrium.

Definition 2 (Non-Equilibrium Structural Response)**.** *The truncated non-stationary diffusion operator and its corresponding node-wise response are defined as*

$$\mathbf{K}_t^{(M)} = \sum_{r=2}^{M+1} e^{-t\mu_r}\mathbf{P}_r, \qquad s_t(u) = \mathbf{e}_u^\top \mathbf{K}_t^{(M)} \mathbf{e}_u. \tag{4}$$

For a simple spectrum, $s_t(u) = \sum_{k=2}^{M+1} e^{-t\lambda_k}\phi_k(u)^2$, which corresponds to the practical NEDP response with $M = m$.

Eq. (4) is a truncated heat-kernel diagonal with the stationary mode removed and is closely related to the Heat Kernel Signature (Sun et al., 2009). It provides a diffusion-weighted structural response that captures transient graph structure rather than a direct foreground probability.

Lemma 1 (Spectral Representation Invariance)**.** *The response $s_t(u)$ is invariant to eigenvector sign and, when complete eigenspaces are retained, to orthogonal basis changes within repeated eigenspaces.*

We consider three oracle conditions: (A1) $C_M \geq 1-\epsilon$, with $0 \leq \epsilon < 1$; (A2) $\mathcal{D}_M \geq \delta$, with $0 < \delta < 1$; and (A3) non-empty mode sets $\mathcal{S}, \mathcal{N} \subseteq \{2, \ldots, M+1\}$ satisfy $\mu_s = \max_{r\in\mathcal{S}} \mu_r < \mu_n = \min_{r\in\mathcal{N}} \mu_r$. These conditions characterize the oracle spectral regime considered in the following analysis and are not used during practical pseudo-label generation.

Theorem 1 (Oracle Finite-Time Multi-Mode Advantage)**.** *Define the Fiedler-only and finite-time multi-mode reconstruction errors as*

$$\mathcal{E}_2 = \frac{\|\bar{\mathbf{g}} - \mathbf{P}_2\bar{\mathbf{g}}\|_2^2}{\|\bar{\mathbf{g}}\|_2^2}, \qquad \mathcal{E}_t^{(M)} = \frac{\|\bar{\mathbf{g}} - \mathbf{K}_t^{(M)}\bar{\mathbf{g}}\|_2^2}{\|\bar{\mathbf{g}}\|_2^2}. \tag{5}$$

Under (A1)–(A2),

$$\mathcal{E}_2 - \mathcal{E}_t^{(M)} \geq \left[\mathcal{D}_M - (1 - e^{-t\mu_{M+1}})^2\right] C_M. \tag{6}$$

Hence, defining $t_{\mathrm{pres}} = -\log(1-\sqrt{\delta})/\mu_{M+1}$, any $0 \leq t < t_{\mathrm{pres}}$ satisfies $\mathcal{E}_t^{(M)} < \mathcal{E}_2$, preserving the multi-mode reconstruction advantage.

The complete derivation is provided in Appendix A.2.1. The theorem concerns the oracle-filtered signal $\mathbf{K}_t^{(M)}\bar{\mathbf{g}}$ and therefore does not imply that the practical response $s_t(u)$ directly estimates the foreground mask.

2.3 Conditional Contrast Amplification and Non-Equilibrium Window

We further analyze the relative attenuation of informative and nuisance spectral groups under finite-time diffusion to characterize an effective non-equilibrium regime.

Theorem 2 (Conditional Spectral Contrast Amplification)**.** *For node u, define*

$$s_t^{\mathcal{S}}(u) = \sum_{r\in\mathcal{S}} e^{-t\mu_r}\mathbf{e}_u^\top \mathbf{P}_r \mathbf{e}_u, \qquad s_t^{\mathcal{N}}(u) = \sum_{r\in\mathcal{N}} e^{-t\mu_r}\mathbf{e}_u^\top \mathbf{P}_r \mathbf{e}_u. \tag{7}$$

Under (A3), whenever $s_0^{\mathcal{N}}(u) > 0$,

$$\frac{s_t^{\mathcal{S}}(u)}{s_t^{\mathcal{N}}(u)} \geq e^{t(\mu_n - \mu_s)} \frac{s_0^{\mathcal{S}}(u)}{s_0^{\mathcal{N}}(u)}. \tag{8}$$

Thus, finite-time diffusion progressively strengthens the relative contribution of the lower-frequency informative group. However, an excessively large t also suppresses all non-stationary responses. The complete proof is given in Appendix A.2.2.

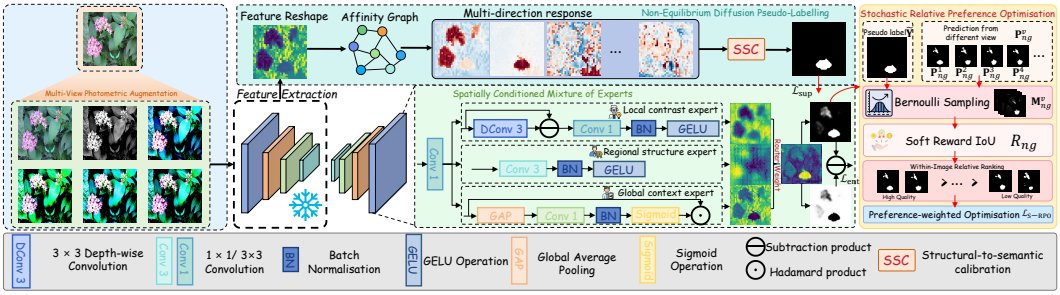


Figure 3: **The main framework of our proposed method.** The model contains a frozen encoder, a non-equilibrium diffusion pseudo-labelling strategy, a spatially conditioned mixture of experts decoder, and a stochastic relative preference Optimisation strategy.

Corollary 1 (Conditional Non-Equilibrium Diffusion Window)**.** *Let* $t_{\mathrm{con}} = \log \Gamma/(\mu_n - \mu_s)$ *and* $t_{\mathrm{ret}} = -\log \alpha/\mu_s$*, where* $\Gamma > 1$ *is the desired contrast gain and* $\alpha \in (0, 1)$ *is the minimum retained informative response. Then a sufficient non-equilibrium diffusion window is*

$$t_{\mathrm{con}} \le t < \min\{t_{\mathrm{ret}}, t_{\mathrm{pres}}\}. \tag{9}$$

When non-empty, this interval balances relative contrast amplification, informative-response retention, and preservation of the oracle multi-mode advantage. These quantities are analytical only; in practice, t is treated as a fixed global hyperparameter and validated empirically.

3 Methodology

Given a mini-batch of unlabeled images $\mathbf{X} \in \mathbb{R}^{B\times 3\times H\times W}$, a frozen encoder extracts features $\mathbf{Z} \in \mathbb{R}^{B\times C\times h\times w}$. For the n-th image, $\mathbf{Z}_n$ is reshaped into $\mathbf{F}_n \in \mathbb{R}^{N\times C}$ with $N = hw$. As shown in Fig. 3, DES-UCOD consists of three components. NEDP operates on $\mathbf{F}_n$ to discover complementary structures and generate pseudo-labels $\widetilde{\mathbf{Y}}_n$. SC-MoE processes $\mathbf{Z}_n$ with spatially Specialised experts to produce foreground and background predictions $\mathbf{O}^f_{ng}$ and $\mathbf{0}^b_{ng}$. Finally, S-RPO performs relative preference Optimisation over aligned augmented views to mitigate label noise. *See Appendix A.3 for the complete training and inference procedure.*

3.1 Non-Equilibrium Diffusion Pseudo-Labelling

Intra-object multi-camouflage may distribute object evidence beyond the dominant Fiedler direction, resulting in incomplete spectral partitions. Motivated by Section 2, NEDP aggregates complementary low-frequency responses before diffusion equilibrium. For the n-th image, we reshape $\mathbf{Z}_n$ into $\mathbf{F}_n$. Omitting n, let $\mathbf{f}_u \in \mathbb{R}^C$ denote the u-th patch feature and $\widehat{\mathbf{f}}_u = \mathbf{f}_u/\|\mathbf{f}_u\|_2$ its normalised representation. We construct the affinity matrix $\mathbf{A}$ as

$$A_{uv} = \begin{cases} \max([\widehat{\mathbf{f}}_u^\top \widehat{\mathbf{f}}_v]_+, \epsilon_a), & u \neq v, \\ 0, & u = v, \end{cases} \tag{10}$$

where $[\cdot]_+ = \max(\cdot, 0)$ and $\epsilon_a = 10^{-6}$. The affinity floor guarantees graph connectivity and positive node degrees. With $d_u = \sum_v A_{uv}$ and $\mathbf{D}^{-1/2} = \mathrm{diag}(d_u^{-1/2})$, the symmetric normalised graph Laplacian is $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$. Let $\{(\lambda_k, \boldsymbol{\phi}_k)\}_{k=1}^N$ denote its ordered eigenpairs, where $0 = \lambda_1 < \lambda_2 \le \cdots \le \lambda_N$.

Finite-time multi-direction response. NEDP combines multiple non-stationary directions through finite-time diffusion:

$$\mathbf{K}_t^{(m)} = \sum_{k=2}^{m+1} e^{-t\lambda_k} \boldsymbol{\phi}_k \boldsymbol{\phi}_k^\top, \qquad s_t(u) = \sum_{k=2}^{m+1} e^{-t\lambda_k} \phi_k(u)^2, \tag{11}$$

where $m = 10$ and $t = 10$. Since the graph is connected, $k = 1$ corresponds to the unique stationary mode and is excluded. Squared coordinates remove sign ambiguity, while $e^{-t\lambda_k}$ suppresses higher-frequency responses. Theorems 1 and 2 motivate this construction under the stated oracle conditions, but do not provide a foreground-recovery guarantee for the practical response.

Structural-to-Semantic Calibration (SSC). Since $s_t(u)$ has no foreground polarity, we normalise it to $\bar{s}_t(u) \in [0,1]$ and reshape it into the $h \times w$ patch grid. Otsu thresholding produces a binary partition, whose polarity is determined by the boundary prior:

$$Q_u = \mathbb{I}[\bar{s}_t(u) > \mathrm{Otsu}(\bar{\mathbf{s}}_t)], \quad b_i = \begin{cases} \sum_{u\in\partial\Omega} \mathbb{I}[Q_u = 0], & i = 0, \\ \sum_{u\in\partial\Omega} \mathbb{I}[Q_u = 1], & i = 1, \end{cases} \quad Q^p = \begin{cases} Q, & b_0 \geq b_1, \\ 1 - Q, & b_1 > b_0. \end{cases} \tag{12}$$

Here, $\partial\Omega$ denotes the grid boundary, b_i is the boundary support of partition i, and the boundary-dominant partition is assigned to background. Let $\{\mathcal{C}_j\}$ denote the connected foreground regions in Q^p. We retain the largest component and upsample it to obtain the full-resolution pseudo-label:

$$\mathcal{C}_{\max} = \arg\max_{\mathcal{C}_j} |\mathcal{C}_j|, \qquad \widetilde{Y}^l_u = \mathbb{I}[u \in \mathcal{C}_{\max}], \qquad \widetilde{\mathbf{Y}} = \mathbb{I}[\mathrm{Up}(\widetilde{\mathbf{Y}}^l) > 0.5], \tag{13}$$

where $|\mathcal{C}_j|$ denotes region area and $\mathrm{Up}(\cdot)$ denotes bilinear interpolation to $H \times W$. Residual pseudo-label errors are subsequently handled by S-RPO.

3.2 Stochastic Relative Preference Optimisation

Although NEDP improves pseudo-label completeness, its residual errors may produce misleading rewards. We therefore introduce S-RPO, which uses detached agreement among aligned views to measure the predictive support for each pseudo-label pixel and reduce the reward contribution of locally inconsistent regions.

Aligned stochastic exploration. For the n-th image $\mathbf{X}_n$, we construct G aligned views $\mathbf{X}_{ng} = \mathcal{T}_g(\mathbf{X}_n)$ for $g = 1, \ldots, G$, with $G = 6$. The transformations modify image appearance while preserving spatial correspondence, so all views share the NEDP pseudo-label $\widetilde{\mathbf{Y}}_n$. Given the foreground prediction $\mathbf{P}_{ng} \in [0,1]^{H\times W}$, we draw a relaxed Bernoulli sample $\mathbf{M}_{ng}$ for each pixel $i \in \Omega$:

$$\mathbf{M}_{ngi} = \sigma\left(\frac{\mathrm{logit}(\mathbf{P}_{ngi}) + \log U_{ngi} - \log(1 - U_{ngi})}{\tau_s}\right), \tag{14}$$

where $U_{ngi} \sim \mathcal{U}(0,1)$ and $\tau_s = 0.5$ is the relaxation temperature.

Cross-view-supported structural reward. Direct comparison with the complete pseudo-label may amplify residual errors. For each view g, we estimate the cross-view support for each pseudo-label pixel using the remaining $G - 1$ aligned predictions:

$$\overline{\mathbf{P}}_{n,\bar{g},i} = \mathrm{sg}\left(\frac{1}{G-1}\sum_{g'\neq g} \mathbf{P}_{ng'i}\right), \qquad c_{ngi} = \mathrm{sg}(1 - |\widetilde{\mathbf{Y}}_{ni} - \overline{\mathbf{P}}_{n,\bar{g},i}|), \tag{15}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient and $c_{ngi} \in [0,1]$. Excluding the current view prevents it from determining its own confidence. A large value of c_{ngi} indicates that the pseudo-label is supported by the remaining views, whereas a small value indicates disagreement. We retain the NEDP pseudo-label as the structural reference and use c_{ngi} to weight its contribution. The confidence-weighted soft IoU is

$$J_{ng} = \frac{\sum_{i\in\Omega} c_{ngi}\mathbf{M}_{ngi}\widetilde{\mathbf{Y}}_{ni}}{\sum_{i\in\Omega} c_{ngi}(\mathbf{M}_{ngi} + \widetilde{\mathbf{Y}}_{ni} - \mathbf{M}_{ngi}\widetilde{\mathbf{Y}}_{ni}) + \epsilon_{\mathrm{num}}}, \tag{16}$$

where ϵ_{num} is a numerical stabiliser. Regions unsupported by cross-view predictions therefore contribute less to the reward. We further define $R_{ng} = J_{ng} + \rho_j$ for $J_{ng} \in \mathcal{B}_j$, where $\mathcal{B}_1 = [0, 0.5]$, $\mathcal{B}_2 = (0.5, 0.7]$, $\mathcal{B}_3 = (0.7, 0.8]$, and $\mathcal{B}_4 = (0.8, 1]$, with $\rho_1 = 0$, $\rho_2 = 1$, $\rho_3 = 2$, and $\rho_4 = 3$. The tiered bonus increases the reward separation between predictions of different structural quality.

Within-image relative preference. To reduce image-specific difficulty bias, rewards are normalised across the G views:

$$\omega_{ng} = \frac{R_{ng} - \min_{g'} R_{ng'}}{\max_{g'} R_{ng'} - \min_{g'} R_{ng'} + \epsilon_{\mathrm{num}}}. \tag{17}$$

The weight $\omega_{ng} \in [0,1]$ favours relatively better predictions. When all rewards are identical, the weights become zero because no relative preference can be established. Let $\widehat{\mathbf{M}}_{ngi} = \mathrm{sg}(\mathbf{M}_{ngi})$ and $\widehat{\omega}_{ng} = \mathrm{sg}(\omega_{ng})$. The S-RPO objective is

$$\mathcal{L}_{\text{S-RPO}} = -\frac{1}{BG|\Omega|}\sum_{n=1}^{B}\sum_{g=1}^{G}\widehat{\omega}_{ng}\sum_{i\in\Omega}\log p_{\tau_s}(\widehat{\mathbf{M}}_{ngi} \mid \mathbf{P}_{ngi}), \tag{18}$$

Table 1: Comparison of our method with recent methods on four datasets, with results reported separately for different backbone networks. The best two results are in **red** and **blue** fonts.

Methods	Publications	*CHAMELEON*				*CAMO*				*COD10K*				*NC4K*			
		$M\downarrow$	$F_\beta^w\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
FreeSOLO (Wang et al., 2022a)	CVPR22	0.101	0.503	0.730	0.659	0.177	0.333	0.600	0.534	0.093	0.421	0.725	0.630	0.128	0.441	0.781	0.611
DiffCut (Couairon et al., 2024)	NeurIPS24	0.245	0.374	0.592	0.537	0.173	0.431	0.701	0.573	0.160	0.379	0.628	0.591	0.134	0.531	0.727	0.671
Backbone of the DINO																	
LOST (Siméoni et al., 2021)	BMVC21	0.141	0.406	0.732	0.581	0.173	0.431	0.701	0.573	0.095	0.438	0.742	0.631	0.128	0.505	0.755	0.634
Spectral (Melas-Kyriazi et al., 2022)	CVPR22	0.178	0.471	0.678	0.622	0.210	0.487	0.673	0.603	0.169	0.391	0.614	0.588	0.143	0.559	0.733	0.679
TokenCut (Wang et al., 2022b)	CVPR22	0.134	0.508	0.742	0.652	0.160	0.506	0.714	0.641	0.100	0.474	0.739	0.661	0.101	0.608	0.802	0.721
MaskCut (Wang et al., 2023)	CVPR23	0.191	0.473	0.653	0.616	0.218	0.456	0.640	0.593	0.189	0.393	0.616	0.594	0.178	0.524	0.693	0.655
FOUND (Siméoni et al., 2023)	CVPR23	0.218	0.381	0.602	0.552	0.277	0.397	0.578	0.526	0.248	0.264	0.487	0.488	0.218	0.410	0.605	0.569
ProMerge (Li & Shin, 2024)	ECCV24	0.087	0.534	0.757	0.712	0.134	0.542	0.732	0.692	0.078	0.504	0.756	0.714	0.081	0.639	0.817	**0.773**
SdalsNet (Shou et al., 2025)	AAAI25	0.081	0.612	0.835	0.724	0.118	**0.601**	**0.799**	**0.697**	0.072	0.525	0.780	0.697	0.085	0.642	0.824	0.739
EASE (Du et al., 2025a)	CVPR25	0.105	0.550	0.765	0.676	0.166	0.563	0.737	0.653	0.109	0.514	0.732	0.673	0.108	0.633	0.790	0.728
RISE (Du et al., 2025b)	ICCV25	0.090	0.508	0.741	0.679	0.132	0.507	0.715	0.664	0.062	0.526	0.795	**0.724**	0.081	0.633	0.818	0.759
UCOD-DPL (Yan et al., 2025)	CVPR25	**0.072**	**0.625**	**0.854**	**0.734**	**0.108**	**0.621**	**0.801**	**0.706**	**0.059**	**0.577**	**0.822**	**0.727**	**0.074**	**0.680**	**0.851**	**0.761**
SAPNet (Liu et al., 2026b)	TIP26	0.116	0.383	0.614	0.581	0.157	0.337	0.567	0.536	0.077	0.401	0.686	0.614	0.126	0.400	0.604	0.582
Ours	-	**0.069**	**0.632**	**0.844**	**0.741**	**0.114**	0.568	0.761	0.680	**0.057**	**0.547**	**0.810**	0.711	**0.071**	**0.670**	**0.846**	0.758
Backbone of the DINOv2																	
TokenCut (Wang et al., 2022b)	CVPR22	0.088	0.494	0.705	0.682	0.144	0.439	0.655	0.620	0.088	0.373	0.649	0.619	0.108	0.492	0.699	0.664
MaskCut (Wang et al., 2023)	CVPR23	0.115	0.531	0.719	0.687	0.194	0.494	0.658	0.630	0.190	0.352	0.572	0.574	0.177	0.503	0.668	0.646
FOUND (Siméoni et al., 2023)	CVPR23	0.104	0.490	0.745	0.659	0.172	0.373	0.622	0.558	0.100	0.386	0.675	0.605	0.131	0.455	0.693	0.621
VoteCut (Arica et al., 2024)	CVPR24	0.145	0.546	0.735	0.674	0.170	0.522	0.702	0.637	0.092	0.534	0.782	0.690	0.097	0.649	0.818	0.739
ProMerge (Li & Shin, 2024)	ECCV24	0.085	0.567	0.787	0.741	0.142	0.502	0.706	0.679	0.081	0.435	0.714	0.674	0.096	0.558	0.771	0.726
EASE (Du et al., 2025a)	CVPR25	0.044	0.741	0.899	0.819	0.098	0.684	0.831	0.749	**0.040**	0.656	0.866	0.773	0.056	0.735	0.884	0.800
RISE (Du et al., 2025b)	ICCV25	0.050	0.720	0.884	0.822	0.109	0.610	0.787	0.734	0.049	0.600	0.840	0.763	0.061	0.705	0.868	0.805
EReCu (Jiang et al., 2026)	CVPR26	0.052	0.759	0.904	0.811	0.101	0.674	0.820	0.735	0.042	**0.698**	**0.886**	**0.788**	0.066	**0.747**	0.877	0.793
DualUCOD (Liu et al., 2026a)	ICML26	**0.041**	0.754	**0.918**	**0.827**	**0.083**	**0.702**	**0.852**	**0.771**	0.042	0.635	0.856	0.761	**0.052**	0.745	**0.896**	**0.808**
Ours	-	**0.036**	**0.796**	**0.927**	**0.851**	**0.081**	**0.723**	**0.855**	**0.774**	**0.035**	**0.696**	**0.887**	**0.793**	**0.047**	**0.783**	**0.911**	**0.827**

where $p_{\tau_s}(\widehat{\mathbf{M}}_{ngi} \mid \mathbf{P}_{ngi})$ denotes the relaxed Bernoulli density induced by Eq. (14). Thus, S-RPO retains NEDP as the structural reference while reducing the reward influence of pseudo-label regions unsupported by aligned predictions.

3.3 Spatially Conditioned Mixture of Experts

Intra-object multi-camouflage induces heterogeneous patterns across object regions, making uniform modelling difficult. To capture these variations, we introduce SC-MoE to exploit local contrast, regional structure, and global context. For each aligned view, the encoder feature $\mathbf{Z}_{ng}$ is projected by a 1×1 convolution to $\mathbf{U}_{ng}$ with half the channel dimension.

Local Contrast Expert (LCE). To highlight subtle local variations around camouflage boundaries, we construct a residual response:

$$\mathbf{E}_{ng}^{\text{loc}} = GELU(BN(Conv_{1\times1}(DWConv_{3\times3}(\mathbf{U}_{ng}) - \mathbf{U}_{ng}))), \tag{19}$$

where $DWConv_{3\times3}(\cdot)$ performs channel-wise local aggregation, $Conv_{1\times1}(\cdot)$ denotes 1×1 convolution, $BN(\cdot)$ is batch normalisation, and $GELU(\cdot)$ is the activation function.

Regional Structure Expert (RSE). To capture spatially separated but coherent object regions, we enlarge the receptive field using dilated convolution:

$$\mathbf{E}_{ng}^{\text{reg}} = GELU(BN(Conv_{3\times3}^{d=2}(\mathbf{U}_{ng}))), \tag{20}$$

where $d = 2$ is the dilation rate, enabling regional aggregation without reducing spatial resolution.

Global Context Expert (GCE). To suppress locally ambiguous background responses, we derive an image-level channel prior:

$$\mathbf{E}_{ng}^{\text{ctx}} = \mathbf{U}_{ng} \odot \sigma(BN(Conv_{1\times1}(GAP(\mathbf{U}_{ng})))), \tag{21}$$

where $\text{GAP}(\cdot)$ denotes global average pooling, $\sigma(\cdot)$ is the sigmoid function, and $\odot$ denotes element-wise multiplication.

Spatial Routing and Fusion. A router predicts spatial logits $r_{ng}^e(u)$ for each expert $e \in \{\text{loc}, \text{reg}, \text{ctx}\}$. The corresponding routing weights $\pi_{ng}^e(u)$ and fused representation $\mathbf{V}_{ng}$ are

$$\pi_{ng}^e(u) = \frac{\exp(r_{ng}^e(u))}{\sum_j \exp(r_{ng}^j(u))}, \qquad \mathbf{V}_{ng} = Conv_{3\times3}(\sum_e \boldsymbol{\pi}_{ng}^e \odot \mathbf{E}_{ng}^e), \tag{22}$$

where $\boldsymbol{\pi}_{ng}^e$ is broadcast along the channel dimension. Finally, two independent 3×3 convolutional heads produce foreground and background logits $\mathbf{O}_{ng}^f$ and $\mathbf{O}_{ng}^b$:

$$\mathbf{O}_{ng}^c = Conv_{3\times3}(\mathbf{V}_{ng}), \quad c \in \{f, b\}, \qquad \mathbf{P}_{ng} = \sigma(\mathbf{O}_{ng}^f - \mathbf{O}_{ng}^b). \tag{23}$$

Thus, SC-MoE spatially selects complementary local, regional, and global cues according to the camouflage characteristics of each region.

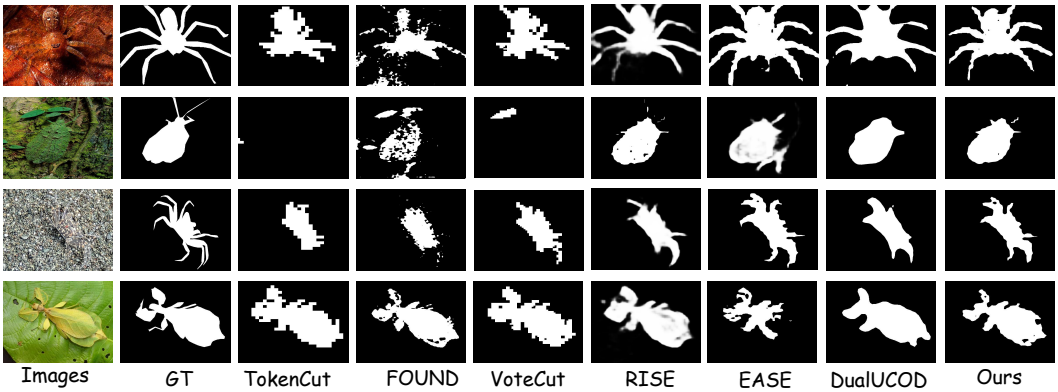


Figure 4: Visual comparison of our method with other approaches. Our method demonstrates a better understanding of camouflaged objects and produces more precise segmentation results.

Table 2: Ablation study on key components.

Model	*COD10K*				*NC4K*			
	$M\downarrow$	$F_\beta^w\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
w/o NEDP	0.052	0.624	0.814	0.752	0.051	0.747	0.882	0.811
w/o S-RPO	0.040	0.677	0.877	0.780	0.058	0.757	0.890	0.805
w/o SC-MoE	0.038	0.685	0.877	0.786	0.052	0.769	0.900	0.816
Ours	**0.035**	**0.696**	**0.887**	**0.793**	**0.047**	**0.783**	**0.911**	**0.827**

Table 3: Analysis of NEDP strategy.

Model	*COD10K*				*NC4K*			
	$M\downarrow$	$F_\beta^w\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
Spectral	0.054	0.622	0.811	0.750	0.053	0.743	0.878	0.810
DESPL	0.040	0.641	0.839	0.761	0.048	0.748	0.888	0.810
w/o SSC	0.039	0.672	0.859	0.780	0.047	0.777	0.903	0.825
Ours	**0.035**	**0.696**	**0.887**	**0.793**	**0.047**	**0.783**	**0.911**	**0.827**

Table 4: Ablation on SC-MoE.

Model	*COD10K*				*NC4K*			
	$M\downarrow$	$F_\beta^w\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
w/o LCE	0.036	0.680	0.875	0.785	0.047	0.777	0.907	0.825
w/o RSE	0.038	0.670	0.866	0.780	**0.046**	0.774	0.906	0.824
w/o GCE	0.039	0.671	0.868	0.779	0.049	0.769	0.903	0.818
Ours	**0.035**	**0.696**	**0.887**	**0.793**	0.047	**0.783**	**0.911**	**0.827**

Table 5: Inference efficiency comparison.

Metric	EASE	RISE	DualUCOD	UCOD-DPL	Ours
Params (M)↓	304.37	331.34	1368.30	**86.68**	308.97
Trainable (M)↓	**0.00**	24.93	1063.93	0.10	4.60
FLOPs (G)↓	416.19	**22.67**	1372.72	307.10	421.20
Time (ms/image)↓	43.72	**8.61**	85.93	54.00	43.98

3.4 Learning Objective

We use the structural loss $\mathcal{L}_{\text{str}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{IoU}}$. The foreground and background branches are supervised by complementary pseudo-labels:

$$\mathcal{L}_{\text{sup}} = \frac{1}{BG}\sum_{n=1}^{B}\sum_{g=1}^{G}[\mathcal{L}_{\text{str}}(\mathbf{O}_{ng}^{f}, \widetilde{\mathbf{Y}}_n) + \mathcal{L}_{\text{str}}(\mathbf{O}_{ng}^{b}, 1-\widetilde{\mathbf{Y}}_n)]. \tag{24}$$

We further incorporate $\mathcal{L}_{\text{S-RPO}}$ from Eq. (18). For numerical stability, the relaxed Bernoulli log-density is clipped as $\bar{\ell}_{ngi} = \text{clip}(\log p_{\tau_s}(\mathbf{M}_{ngi} \mid \mathbf{P}_{ngi}), -5, 0)$. To prevent premature collapse of the stochastic prediction distribution, we introduce

$$\mathcal{L}_{\text{ent}} = \frac{1}{BG|\Omega|}\sum_{n=1}^{B}\sum_{g=1}^{G}\sum_{i\in\Omega}[\mathbf{P}_{ngi}\log\mathbf{P}_{ngi} + (1-\mathbf{P}_{ngi})\log(1-\mathbf{P}_{ngi})]. \tag{25}$$

The final objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{S-RPO}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}}, \tag{26}$$

where $\lambda_{\text{ent}} = 0.1$ controls the entropy regularization strength.

4 Experiments

4.1 Comparison with the state-of-the-art

We compare our method with representative unsupervised approaches, including FreeSOLO (Wang et al., 2022a), DiffCut (Couairon et al., 2024), LOST (Siméoni et al., 2021), Spectral (Melas-Kyriazi et al., 2022), TokenCut (Wang et al., 2022b), MaskCut (Wang et al., 2023), FOUND (Siméoni et al., 2023), VoteCut (Arica et al., 2024), ProMerge (Li & Shin, 2024), SdalsNet (Shou et al., 2025), EASE (Du et al., 2025a), RISE (Du et al., 2025b), EReCu (Jiang et al., 2026), UCOD-DPL (Yan et al., 2025), and DualUCOD (Liu et al., 2026a). All competing methods are evaluated following the experimental settings and data configurations reported in their original papers. *Implementation details, datasets, and evaluation metrics are provided in Appendix A.4.1.*

Quantitative analysis. As shown in Table 1, our DINOv2-based model achieves the best overall performance across all four benchmarks. Compared with DualUCOD on CHAMELEON, it achieves $M = 0.036$, $F_\beta^w = 0.796$, $E_\phi = 0.927$, and $S_\alpha = 0.851$, while also delivering clear improvements

Table 6: Comparison with representative COD methods across supervision settings. F, W, and U denote fully, weakly, and unsupervised learning. "–" indicates unavailable metrics.

Methods	Pub.	Sup.	*CHAMELEON*				*CAMO*				*COD10K*				*NC4K*			
			$M\downarrow$	$F_\beta^w\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
SINet Fan et al. (2020a)	CVPR20	F	0.034	0.806	0.936	0.872	0.092	0.644	0.804	0.745	0.043	0.631	0.864	0.776	0.058	0.723	0.871	0.808
FSPNet Huang et al. (2023)	CVPR23	F	0.023	0.851	0.943	0.908	0.050	0.799	0.899	0.856	0.026	0.735	0.895	0.851	0.035	0.816	0.915	0.879
CRNet He et al. (2023c)	AAAI23	W	0.046	0.744	0.897	0.818	0.092	0.641	0.815	0.735	0.049	0.576	0.832	0.733	0.063	0.688	0.855	0.775
WS-SAM He et al. (2023b)	NeurIPS23	W	0.046	–	0.897	0.824	0.092	–	0.818	0.759	0.038	–	0.878	0.803	0.052	–	0.886	0.829
Just-a-Hint Chen et al. (2024a)	ECCV24	W	–	–	–	–	0.074	0.727	0.872	0.798	0.042	0.650	0.859	0.784	0.051	0.748	0.889	0.822
SEE He et al. (2025a)	TPAMI25	W	0.044	–	0.903	0.826	0.090	–	0.826	0.765	0.036	–	0.883	0.807	0.051	–	0.891	0.836
SCALER He et al. (2026a)	ECCV26	W	0.042	–	0.912	0.838	0.089	–	0.834	0.778	0.034	–	0.882	0.815	0.050	–	0.897	0.817
Ours	–	U	**0.036**	**0.796**	**0.927**	**0.851**	**0.081**	**0.723**	**0.855**	**0.774**	**0.035**	**0.696**	**0.887**	**0.793**	**0.047**	**0.783**	**0.911**	**0.827**

Table 7: Parameter analysis of m, t, G, τ_s, and λ_{ent} on COD10K.

Metrics	m				t				G				τ_s				λ_{ent}			
	6	8	10	12	5	10	15	20	2	4	6	8	0.3	0.5	0.7	1.0	0.05	0.1	0.2	0.3
$M\downarrow$	0.039	0.037	**0.035**	0.037	0.036	**0.035**	0.039	0.037	0.040	0.036	**0.035**	0.040	0.037	**0.035**	0.039	0.038	0.038	**0.035**	0.038	0.038
$F_\beta^w\uparrow$	0.675	0.674	**0.696**	0.677	0.683	**0.696**	0.681	0.684	0.675	0.689	**0.696**	0.672	0.684	**0.696**	0.668	0.682	0.678	**0.696**	0.674	0.673
$E_\phi\uparrow$	0.870	0.868	**0.887**	0.872	0.872	**0.887**	0.870	0.876	0.868	0.876	**0.887**	0.861	0.872	**0.887**	0.857	0.868	0.869	**0.887**	0.865	0.864
$S_\alpha\uparrow$	0.782	0.782	**0.793**	0.783	0.787	**0.793**	0.787	0.787	0.783	0.790	**0.793**	0.779	0.788	**0.793**	0.778	0.787	0.783	**0.793**	0.782	0.781

on CAMO and NC4K. On COD10K, it reduces M to 0.035 and improves F_β^w, E_ϕ, and S_α over EASE by 0.040, 0.021, and 0.020, respectively. These consistent gains demonstrate the effectiveness of NEDP, S-RPO, and SC-MoE in complementary structural recovery, noise-robust learning, and adaptive decoding. With DINO, our method remains competitive and outperforms SAPNet on CHAMELEON, while the remaining gaps on CAMO, COD10K, and NC4K mainly result from the weaker structural representation and feature separability of DINO compared with DINOv2.

Qualitative analysis. As shown in Fig. 4, existing methods often produce incomplete masks or false responses under complex camouflage. In contrast, our method recovers more complete objects with cleaner boundaries, benefiting from NEDP, S-RPO, and SC-MoE.

4.2 Ablation Study and Analysis

We conduct ablation experiments on COD10K-Test and NC4K to validate the effectiveness and contributions of the proposed components. *See Appendix A.4.2 for extended experimental analysis.*

Component Ablation. As shown in Table 2, removing NEDP, S-RPO, or SC-MoE generally degrades performance, confirming their complementary roles. NEDP is particularly important on COD10K, where its removal increases M from 0.035 to 0.052 and decreases F_β^w from 0.696 to 0.624. Table 3 further shows clear gains over Spectral and DESPL. Removing SSC reduces F_β^w, E_ϕ, and S_α on both datasets, while M remains 0.047 on NC4K, indicating its stronger contribution to structural quality. Removing S-RPO increases M to 0.040 and 0.058 on COD10K and NC4K, respectively, confirming its effectiveness against residual pseudo-label noise. Table 4 also validates the complementary roles of LCE, RSE, and GCE. Although removing RSE slightly improves M on NC4K from 0.047 to 0.046, the other metrics decrease, indicating no overall improvement.

Further Analysis. Table 5 shows that DES-UCOD uses only 4.60M trainable parameters and runs at 43.98 ms/image, comparable to EASE and substantially more efficient than DualUCOD. Although RISE is faster, it achieves notably lower segmentation accuracy, indicating a reasonable accuracy-efficiency trade-off. As shown in Table 6, despite using no human annotations, our method achieves lower M and higher E_ϕ than all listed weakly supervised methods on CHAMELEON and NC4K. Remaining gaps on CAMO and COD10K are reasonable since weakly supervised methods exploit human-provided cues. Finally, Table 7 shows that $m = 10$, $t = 10$, $G = 6$, $\tau_s = 0.5$, and $\lambda_{\text{ent}} = 0.1$ provide the best overall performance, supporting a balanced configuration for structural diffusion and stochastic exploration.

5 Conclusion

In this work, we propose DES-UCOD, a Discover-Explore-Specialise framework for unsupervised camouflaged object detection. NEDP uses non-equilibrium diffusion to integrate complementary cues and generate more complete pseudo-labels under intra-object multi-camouflage. S-RPO mitigates residual pseudo-label noise through reward-guided relative preference learning, while SC-MoE models heterogeneous camouflage patterns with spatially Specialised experts. Extensive experiments on four benchmarks demonstrate the effectiveness of DES-UCOD over existing methods.

AI USE STATEMENT

Generative AI tools were used during manuscript revision to assist with language editing, structural organisation, and critical feedback on clarity and presentation. These tools were not treated as sources of scientific evidence and did not generate or alter the experimental data or reported results. All mathematical statements, methodological descriptions, citations, and empirical claims included in the manuscript were independently examined and verified by the authors. The authors made all final decisions and take full responsibility for the content of this work.

REFERENCES

Shahaf Arica, Or Rubin, Sapir Gershov, and Shlomi Laufer. Cuvler: Enhanced unsupervised object discoveries through exhaustive self-supervised transformers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23105–23114. IEEE, 2024.

Huafeng Chen, Dian Shao, Guangqian Guo, and Shan Gao. Just a hint: Point-supervised camouflaged object detection. In *European Conference on Computer Vision*, pp. 332–348. Springer, 2024a.

Zhongxi Chen, Ke Sun, and Xianming Lin. CamoDiffusion: Camouflaged object detection via conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1272–1280, 2024b. doi: 10.1609/aaai.v38i2.27890.

Paul Couairon, Mustafa Shukor, Jean-Emmanuel Haugeard, Matthieu Cord, and Nicolas Thome. Diffcut: Catalyzing zero-shot semantic segmentation with diffusion features and recursive normalized cut. *Advances in Neural Information Processing Systems*, 37:13548–13578, 2024.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

Ji Du, Fangwei Hao, Mingyang Yu, Desheng Kong, Jiesheng Wu, Bin Wang, Jing Xu, and Ping Li. Shift the lens: Environment-aware unsupervised camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19271–19282, 2025a.

Ji Du, Xin Wang, Fangwei Hao, Mingyang Yu, Chunyuan Chen, Jiesheng Wu, Bin Wang, Jing Xu, and Ping Li. Beyond single images: Retrieval self-augmented unsupervised camouflaged object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22131–22142, 2025b.

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.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 *International conference on medical image computing and computer-assisted intervention*, pp. 263–273. Springer, 2020b.

Fen Fang, Liyuan Li, Ying Gu, Hongyuan Zhu, and Joo-Hwee Lim. A novel hybrid approach for crack detection. *Pattern Recognition*, 107:107474, 2020.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22046–22055, 2023a. doi: 10.1109/CVPR52729.2023.02111.

Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with SAM-based pseudo labeling and multi-scale feature grouping. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Zhenhua Guo, Xiu Li, Martin Danelljan, and Fisher Yu. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. In *International Conference on Learning Representations*, 2024.

Chunming He, Kai Li, Yachao Zhang, Ziyun Yang, Youwei Pang, Longxiang Tang, Chengyu Fang, Yulun Zhang, Linghe Kong, Xiu Li, and Sina Farsiu. Segment concealed objects with incomplete supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a. doi: 10.1109/TPAMI.2025.3576209.

Chunming He, Rihan Zhang, Fengyang Xiao, Chengyu Fang, Longxiang Tang, Yulun Zhang, Linghe Kong, Deng-Ping Fan, Kai Li, and Sina Farsiu. RUN: Reversible unfolding network for concealed object segmentation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22853–22864. PMLR, 2025b.

Chunming He, Rihan Zhang, Longxiang Tang, Ziyun Yang, Kai Li, Deng-Ping Fan, and Sina Farsiu. When specialists meet generalists: Segmenter-coordinated asymmetric learning for label-deficient concealed object segmentation. In *European Conference on Computer Vision*, 2026a.

Chunming He, Rihan Zhang, Fengyang Xiao, Dingming Zhang, Zhiwen Cao, and Sina Farsiu. Refining context-entangled content segmentation via curriculum selection and anti-curriculum promotion. In *International Conference on Machine Learning*, 2026b.

Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson W. H. Lau. Weakly-supervised camouflaged object detection with scribble annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 781–789, 2023c. doi: 10.1609/aaai.v37i1.25156.

Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5557–5566, 2023. doi: 10.1109/CVPR52729.2023.00538.

Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4713–4722, 2022.

Shuo Jiang, Gaojia Zhang, Min Tan, Yufei Yin, and Gang Pan. EReCu: Pseudo-label evolution fusion and refinement with multi-cue learning for unsupervised camouflage detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25547–25556, June 2026.

Nicolas Keriven. Not too little, not too much: a theoretical analysis of graph (over)smoothing. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer vision and image understanding*, 184:45–56, 2019.

Minho Lee, Minwoo Choi, Taehoon Yang, Jingu Kim, Jaeoh Kim, Ojeong Kwon, and Namsuk Cho. A study on the advancement of intelligent military drones: Focusing on reconnaissance operations. *IEEE Access*, 12:55964–55975, 2024.

Dylan Li and Gyungin Shin. Promerge: Prompt and merge for unsupervised instance segmentation. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Pingzhu Liu, Chunming He, Zunnan Xu, Chao Hao, Bo Zhao, Xingyu Shao, Jun Zhou, Zitong Yu, and Xiu Li. Unsupervised Camouflaged Object Detection with Dual-Eigenvector Spectral Pseudo-Labeling and Contrastive Refinement. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2026a.

Shijie Liu, Binwei Xu, Tuo Shen, Guanghui Yue, and Qiuping Jiang. Self-anchored progressive framework with noise mitigation for unsupervised camouflaged object detection. *IEEE Transactions on Image Processing*, 35:3581–3593, 2026b.

Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11591–11601, 2021.

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.

Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8364–8375, 2022.

Boaz Nadler and Meirav Galun. Fundamental limitations of spectral clustering. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

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

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2160–2170, 2022.

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.

Peiyao Shou, Yixiu Liu, Wei Wang, Yaoqi Sun, Zhigao Zheng, Shangdong Zhu, and Chenggang Yan. Sdalsnet: Self-distilled attention localization and shift network for unsupervised camouflaged object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6914–6921, 2025.

Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *British Machine Vision Conference*, 2021.

Oriane Siméoni, Chloé Sekkat, Gilles Puy, Antonín Vobecký, Éloi Zablocki, and Patrick Pérez. Unsupervised object localization: Observing the background to discover objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3176–3186, 2023.

Przemysław Skurowski, Hassan Abdulameer, Jakub Błaszczyk, Tomasz Depta, Adam Kornacki, and Przemysław Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6):7, 2018.

Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. *Computer Graphics Forum*, 28(5):1383–1392, 2009. doi: 10.1111/j.1467-8659.2009.01515.x.

Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M. Alvarez. FreeSOLO: Learning to segment objects without annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a.

Xudong Wang, Rohit Girdhar, Stella X. Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3124–3134, 2023.

Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14543–14553, 2022b.

Fengyang Xiao, Sujie Hu, Yuqi Shen, Chengyu Fang, Jinfa Huang, Chunming He, Longxiang Tang, Ziyun Yang, and Xiu Li. A survey of camouflaged object detection and beyond. *arXiv preprint arXiv:2408.14562*, 2024a.

Fengyang Xiao, Pan Zhang, Chunming He, Runze Hu, and Yutao Liu. Concealed object segmentation with hierarchical coherence modeling. In *Artificial Intelligence – CICAI 2023*, volume 14473 of *Lecture Notes in Computer Science*, pp. 16–27. Springer, 2024b. doi: 10.1007/978-981-99-8850-1_2.

Weiqi Yan, Lvhai Chen, Huaijia Kou, Shengchuan Zhang, Yan Zhang, and Liujuan Cao. Ucod-dpl: Unsupervised camouflaged object detection via dynamic pseudo-label learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30365–30375, 2025.

Yi Zhang and Chengyi Wu. Unsupervised camouflaged object segmentation as domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4334–4344, 2023.

A Appendix

A.1 Related Work

A.1.1 Camouflaged Object Detection

Research in camouflaged object detection has progressively shifted from fully supervised paradigms toward settings requiring fewer or no annotations (Xiao et al., 2024a). Fully supervised methods rely on dense labels to capture camouflage-specific cues, employing strategies like search-identification (Fan et al., 2020a; Mei et al., 2021), multi-scale refinement (Pang et al., 2022; Jia et al., 2022; Huang et al., 2023), diffusion modelling (Chen et al., 2024b), and structural coherence modelling (He et al., 2023a; Xiao et al., 2024b; He et al., 2024; 2025b; 2026b). To reduce labeling costs, weakly supervised approaches exploit scribble annotations (He et al., 2023c) or SAM-assisted pseudo-labelling under sparse supervision (He et al., 2023b; 2025a). However, their inherent dependence on human-provided labels motivates the exploration of unsupervised alternatives. In this annotation-free setting, UCOD-DPL (Yan et al., 2025) adaptively combines heuristic pseudo-labels with teacher predictions, while DualUCOD (Liu et al., 2026a) integrates dual-eigenvector pseudo-labelling with contrastive refinement. Following this trajectory, our DES-UCOD explicitly addresses pseudo-label completeness and robust learning against residual labeling errors.

A.1.2 Unsupervised Object Segmentation

Unsupervised object discovery and segmentation exploit pre-trained self-supervised representations without requiring manual annotations. LOST (Siméoni et al., 2021) localises objects through seed expansion, while TokenCut (Wang et al., 2022b) and Deep Spectral Methods (Melas-Kyriazi et al., 2022) partition feature graphs using Laplacian eigenvectors. FOUND (Siméoni et al., 2023) identifies the foreground through background estimation and self-supervised refinement. FreeSOLO (Wang et al., 2022a) and CutLER (Wang et al., 2023) train instance segmenters using automatically generated masks. Despite their success on general-object and saliency benchmarks, their direct application to COD is severely hindered by extreme foreground-background similarity and indistinct boundaries (Zhang & Wu, 2023). We further consider the intrinsic challenge of *intra-object multi-camouflage*, where differently textured object parts seamlessly blend into their respective local surroundings. To this end, our NEDP strategy leverages non-equilibrium diffusion to integrate complementary structural cues derived from transient graph responses, aiming to generate structurally complete pseudo-labels across heterogeneous object parts.

A.2 Theoretical Derivations

A.2.1 Proof of Theorem 1

Proof. Recall that $\bar{\mathbf{g}} = (\mathbf{I} - \mathbf{P}_1)\mathbf{g}$ contains no stationary component. Therefore, it admits the spectral decomposition

$$\bar{\mathbf{g}} = \sum_{r=2}^{Q} \mathbf{P}_r \bar{\mathbf{g}}.$$

Since the eigenspace projectors are mutually orthogonal, the retained and discarded spectral components can be analysed independently.

Applying the finite-time diffusion operator $\mathbf{K}_t^{(M)}$ to $\bar{\mathbf{g}}$ gives

$$\mathbf{K}_t^{(M)} \bar{\mathbf{g}} = \sum_{r=2}^{M+1} e^{-t\mu_r} \mathbf{P}_r \bar{\mathbf{g}}.$$

Hence, the reconstruction residual can be decomposed into two parts: the attenuation error within the retained subspace and the energy lying outside that subspace,

$$\bar{\mathbf{g}} - \mathbf{K}_t^{(M)}\bar{\mathbf{g}} = \sum_{r=2}^{M+1} \left(1 - e^{-t\mu_r}\right) \mathbf{P}_r \bar{\mathbf{g}} + \sum_{r=M+2}^{Q} \mathbf{P}_r \bar{\mathbf{g}}. \tag{27}$$

Because these spectral components are mutually orthogonal, their squared norms add without cross terms. Therefore,

$$\begin{aligned} \mathcal{E}_t^{(M)} &= \frac{\left\| \bar{\mathbf{g}} - \mathbf{K}_t^{(M)}\bar{\mathbf{g}} \right\|_2^2}{\|\bar{\mathbf{g}}\|_2^2} \\ &= 1 - C_M + \frac{\sum_{r=2}^{M+1} \left(1 - e^{-t\mu_r}\right)^2 \|\mathbf{P}_r \bar{\mathbf{g}}\|_2^2}{\|\bar{\mathbf{g}}\|_2^2}. \end{aligned} \tag{28}$$

Here, the term $1 - C_M$ corresponds to the foreground-associated spectral energy outside the retained low-frequency subspace, while the second term quantifies the additional attenuation introduced by finite-time diffusion within the retained modes.

For $2 \le r \le M+1$, the ordered spectrum satisfies $\mu_r \le \mu_{M+1}$. Since $1 - e^{-t\mu}$ is monotonically increasing in μ for $t \ge 0$, we have

$$1 - e^{-t\mu_r} \le 1 - e^{-t\mu_{M+1}}.$$

Thus, the attenuation of every retained mode is upper-bounded by that of the highest-frequency retained mode. Using the definition of C_M gives

$$\mathcal{E}_t^{(M)} \le 1 - C_M + C_M \left(1 - e^{-t\mu_{M+1}}\right)^2. \tag{29}$$

We next relate this finite-time error to the Fiedler-only reconstruction. From Definition 1,

$$\mathcal{D}_M = 1 - \frac{\|\mathbf{P}_2 \bar{\mathbf{g}}\|_2^2}{\|\mathbf{\Pi}_M \bar{\mathbf{g}}\|_2^2}.$$

Since $\|\mathbf{\Pi}_M \bar{\mathbf{g}}\|_2^2 = C_M \|\bar{\mathbf{g}}\|_2^2$, we obtain

$$\frac{\|\mathbf{P}_2 \bar{\mathbf{g}}\|_2^2}{\|\bar{\mathbf{g}}\|_2^2} = (1 - \mathcal{D}_M) C_M.$$

Consequently, the Fiedler-only reconstruction error is

$$\mathcal{E}_2 = 1 - C_M + \mathcal{D}_M C_M. \tag{30}$$

The additional term $\mathcal{D}_M C_M$ represents the retained low-frequency energy that is not captured by the Fiedler eigenspace.

Combining Eqs. (29) and (30) yields

$$\mathcal{E}_2 - \mathcal{E}_t^{(M)} \ge \left[\mathcal{D}_M - \left(1 - e^{-t\mu_{M+1}}\right)^2\right] C_M. \tag{31}$$

This inequality makes the trade-off explicit: the multi-mode advantage $\mathcal{D}_M$ must dominate the attenuation introduced by finite-time diffusion.

Under (A1)–(A2), $C_M \ge 1 - \epsilon$ and $\mathcal{D}_M \ge \delta$. Therefore,

$$\mathcal{E}_2 - \mathcal{E}_t^{(M)} \ge \left[\delta - \left(1 - e^{-t\mu_{M+1}}\right)^2\right] (1 - \epsilon).$$

To ensure a strict reconstruction advantage, it is sufficient that

$$\left(1 - e^{-t\mu_{M+1}}\right)^2 < \delta.$$

Since $0 < \delta < 1$, this condition is equivalent to

$$1 - e^{-t\mu_{M+1}} < \sqrt{\delta},$$

or

$$e^{-t\mu_{M+1}} > 1 - \sqrt{\delta}.$$

Taking logarithms gives

$$t < t_{\text{pres}} = -\frac{\log(1-\sqrt{\delta})}{\mu_{M+1}}.$$

Hence, for any $0 \le t < t_{\text{pres}}$,

$$\delta - \left(1 - e^{-t\mu_{M+1}}\right)^2 > 0,$$

and therefore

$$\mathcal{E}_2 - \mathcal{E}_t^{(M)} \ge \left[\delta - \left(1 - e^{-t\mu_{M+1}}\right)^2\right](1-\epsilon) > 0.$$

Thus, sufficiently early diffusion preserves a strict oracle reconstruction advantage of the retained multi-mode representation over the Fiedler-only projection. This proves Theorem 1. □

A.2.2 Proof of Theorem 2 and Corollary 1

Proof. We first establish the relative attenuation of the informative and nuisance spectral groups. Since each $\mathbf{P}_r$ is an orthogonal projector, it is positive semidefinite. Hence, for every node u,

$$\mathbf{e}_u^\top \mathbf{P}_r \mathbf{e}_u \ge 0.$$

This non-negativity allows the exponential spectral weights to be bounded term by term.

For $r \in \mathcal{S}$, Assumption (A3) gives $\mu_r \le \mu_s$. Since $e^{-t\mu}$ decreases with μ for $t \ge 0$,

$$e^{-t\mu_r} \ge e^{-t\mu_s}.$$

Therefore, every informative-mode contribution is attenuated by no more than $e^{-t\mu_s}$, yielding

$$s_t^{\mathcal{S}}(u) \ge e^{-t\mu_s} s_0^{\mathcal{S}}(u). \tag{32}$$

Similarly, for $r \in \mathcal{N}$, Assumption (A3) gives $\mu_r \ge \mu_n$, and hence

$$e^{-t\mu_r} \le e^{-t\mu_n}.$$

The nuisance response therefore satisfies

$$s_t^{\mathcal{N}}(u) \le e^{-t\mu_n} s_0^{\mathcal{N}}(u). \tag{33}$$

Whenever $s_0^{\mathcal{N}}(u) > 0$, the nuisance response remains positive for finite t, because it is a positively weighted sum of non-negative terms. We may therefore divide the two bounds above to obtain

$$\frac{s_t^{\mathcal{S}}(u)}{s_t^{\mathcal{N}}(u)} \ge e^{t(\mu_n - \mu_s)} \frac{s_0^{\mathcal{S}}(u)}{s_0^{\mathcal{N}}(u)}.$$

Since $\mu_n > \mu_s$, the multiplicative factor $e^{t(\mu_n-\mu_s)}$ increases with t. Thus, finite-time diffusion progressively increases the relative contribution of the prescribed lower-frequency informative group. This proves Theorem 2.

We next derive the diffusion-time conditions used in Corollary 1. To obtain a relative contrast gain of at least $\Gamma > 1$, it is sufficient to require

$$e^{t(\mu_n-\mu_s)} \ge \Gamma.$$

Solving for t gives the lower bound

$$t \ge t_{\text{con}} = \frac{\log \Gamma}{\mu_n - \mu_s}.$$

Thus, t_{con} is the minimum diffusion time sufficient to achieve the prescribed relative contrast amplification.

However, increasing t also attenuates the informative response itself. From Eq. (32), retaining at least an $\alpha \in (0, 1)$ fraction of the initial informative response is guaranteed if

$$e^{-t\mu_s} \geq \alpha.$$

This gives the upper bound

$$t \leq t_{\text{ret}} = -\frac{\log \alpha}{\mu_s}.$$

Hence, t_{ret} prevents excessive attenuation of the informative spectral group.

Finally, Theorem 1 requires

$$t < t_{\text{pres}}$$

to preserve the oracle finite-time multi-mode reconstruction advantage. Combining the contrast, retention, and preservation requirements yields the sufficient interval

$$t_{\text{con}} \leq t < \min\{t_{\text{ret}}, t_{\text{pres}}\}.$$

When this interval is non-empty, finite-time diffusion simultaneously provides the desired relative contrast gain, retains sufficient informative response, and preserves the oracle multi-mode advantage. This proves Corollary 1. □

A.2.3 Empirical Evidence for the Multi-Mode Spectral Analysis

The preceding proofs establish a conditional advantage of retaining complementary non-stationary modes. We now examine the empirical relevance of this analysis: whether foreground-associated energy extends beyond the Fiedler eigenspace, whether spectral dispersion is associated with the practical benefit of multi-mode aggregation, and how this benefit changes with the number of retained modes. These experiments connect the spectral quantities in Definition 1 to observable pseudo-mask behaviour.

Evaluation protocol and spectral diagnostics. We analyse the saved outputs of our label-free spectral diagnostic pipeline on CAMO-Train and COD10K-Train. Image features determine the affinity graph and its eigendecomposition. Ground-truth masks are accessed only for post-hoc spectral measurements and evaluation; they are not supplied to graph construction or pseudo-mask generation. Of the 4,040 training images, 4,026 have valid foreground masks after projection onto the 34×34 patch grid: 1,000 from CAMO and 3,026 from COD10K. The remaining 14 masks become empty at this resolution and are excluded from the spectral diagnostics. Following Section 2, let $\mathbf{g} = \mathbf{D}^{1/2}\mathbf{y}$, $\bar{\mathbf{g}} = (\mathbf{I} - \mathbf{P}_1)\mathbf{g}$, and $\mathbf{\Pi}_M = \sum_{r=2}^{M+1} \mathbf{P}_r$. We compute the coverage and Fiedler-relative dispersion for each image as

$$C_M = \frac{\|\mathbf{\Pi}_M \bar{\mathbf{g}}\|_2^2}{\|\bar{\mathbf{g}}\|_2^2}, \qquad \mathcal{D}_M = 1 - \frac{\|\mathbf{P}_2 \bar{\mathbf{g}}\|_2^2}{\|\mathbf{\Pi}_M \bar{\mathbf{g}}\|_2^2} = 1 - \frac{C_1}{C_M}. \tag{34}$$

The additional foreground coverage beyond the Fiedler eigenspace is

$$G_M = \frac{\|(\mathbf{\Pi}_M - \mathbf{P}_2)\bar{\mathbf{g}}\|_2^2}{\|\bar{\mathbf{g}}\|_2^2} = C_M \mathcal{D}_M = C_M - C_1. \tag{35}$$

All reported means are computed from per-image quantities; in particular, $\overline{\mathcal{D}}_M$ is the mean of $1 - C_1/C_M$, rather than a ratio of dataset means.

Foreground energy beyond the Fiedler eigenspace. Figure 5(a) compares C_1 and C_{10}. On the combined set, mean coverage increases from 0.4339 to 0.6588, giving $\overline{G}_{10} = 0.2249$. Thus, the nine additional retained directions increase the captured fraction of total non-stationary foreground energy by 22.49 percentage points on average. The result holds on both datasets: coverage increases from 0.5429 to 0.7212 on CAMO and from 0.3978 to 0.6381 on COD10K. The corresponding additional coverage is 0.1782 and 0.2403, respectively. The mean dispersion is $\overline{\mathcal{D}}_{10} = 0.4153$, indicating that a substantial fraction of the retained energy lies outside the Fiedler eigenspace. To illustrate how often coverage and dispersion occur jointly, 48.7% of valid images satisfy both $C_{10} \geq 0.5$ and $\mathcal{D}_{10} \geq 0.1$. These thresholds are descriptive and are not used for model training or parameter selection. Together, the measurements show that the multi-mode regime is present in a substantial portion of the training data.

Connection to the oracle reconstruction analysis. The quantity G_M has an exact interpretation in the preceding proof. At $t = 0$, $\mathbf{K}_0^{(M)} = \mathbf{\Pi}_M$. Orthogonality of the spectral projectors gives

$$\mathcal{E}_2 - \mathcal{E}_0^{(M)} = C_M - C_1 = G_M. \tag{36}$$

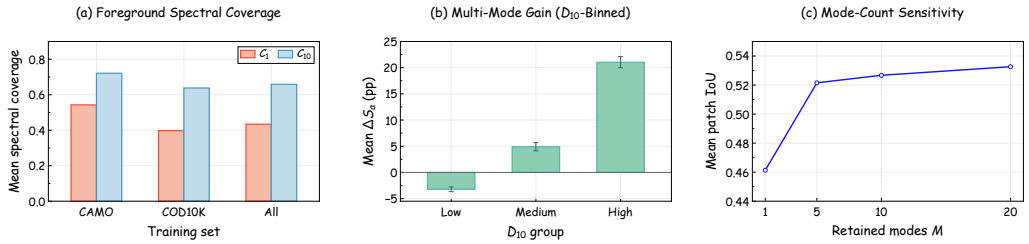


Figure 5: **Empirical evidence for the multi-mode spectral analysis.** (a) Mean coverage of the Fiedler eigenspace (C_1) and the first ten non-stationary modes (C_{10}) on CAMO-Train, COD10K-Train, and the combined set. Their difference is the additional foreground coverage G_{10}, equal to the oracle reconstruction advantage at $t = 0$. (b) Mean full-resolution structural similarity gain, $S_\alpha^{(10)} - S_\alpha^{(\mathrm{Fiedler})}$, in three equal-sized $\mathcal{D}_{10}$ groups (1,342 images each). Gains are multiplied by 100 and expressed in percentage points (pp), not relative percentage changes. Error bars are 95% bootstrap confidence intervals from 10,000 within-group resamples. (c) Mean patch-level IoU for $M = m \in \{1, 5, 10, 20\}$ using the same non-negative response construction. Points are measured settings; lines are visual guides. Increasing from ten to twenty modes adds 0.0059 IoU; ten modes capture 91.7% of the total improvement from one to twenty. All diagnostics use the same 4,026 valid training images, with ground truth used only for post-hoc measurement.

Hence, $\overline{G}_{10} = 0.2249$ is also the mean oracle reconstruction advantage of the ten-mode projector at this endpoint. For finite diffusion time, Theorem 1 bounds the preserved advantage by

$$\mathcal{E}_2 - \mathcal{E}_t^{(M)} \geq C_M \left[\mathcal{D}_M - \left(1 - e^{-t\mu_{M+1}}\right)^2\right]. \tag{37}$$

This expression separates the foreground energy available beyond the Fiedler direction from its attenuation during diffusion. With coverage and attenuation held fixed, greater dispersion increases the lower bound. The observed positive G_{10} therefore quantifies the representational opportunity underlying Theorem 1. We next examine whether a corresponding conditional trend appears in practical pseudo-mask quality.

Dispersion-dependent gains in pseudo-mask quality. We sort all valid images by $\mathcal{D}_{10}$ and divide them into three equal-sized groups, each containing 1,342 images. Their mean dispersions are 0.0544, 0.2662, and 0.9252. We compare the full-resolution ten-mode pseudo-mask with the signed Fiedler partition. For each similarity metric, the gain is $\Delta Q = Q_{10} - Q_{\mathrm{Fiedler}}$; for MAE, we use $\Delta\mathrm{MAE}_{\mathrm{imp}} = \mathrm{MAE}_{\mathrm{Fiedler}} - \mathrm{MAE}_{10}$, so that positive values always favour ten-mode aggregation. Figure 5(b) shows the structural similarity gain, ΔS_α, expressed in percentage points. It changes from -3.21 in the low-dispersion group to $+4.89$ and $+21.03$ in the medium- and high-dispersion groups. Error bars indicate 95% bootstrap confidence intervals, obtained from 10,000 resamples within each fixed group. Table 8 shows that the same group-level pattern appears across all four metrics. In the low-dispersion group, little retained foreground energy is missing from the

Table 8: Ten-mode gains over the signed Fiedler partition for three equally sized dispersion groups. Positive values indicate improvement. Values use the original metric scale; the S_α gains in Figure 5(b) are multiplied by 100.

Group	Mean $\mathcal{D}_{10}$	ΔF_β^w	ΔS_α	ΔE_ϕ	$\Delta\mathrm{MAE}_{\mathrm{imp}}$
Low	0.0544	-0.0229	-0.0321	-0.0293	-0.0191
Medium	0.2662	$+0.0819$	$+0.0489$	$+0.0853$	$+0.0334$
High	0.9252	$+0.2287$	$+0.2103$	$+0.3021$	$+0.2552$

Fiedler representation, and including further modes does not improve the average mask. In the high-dispersion group, ten-mode aggregation improves F_β^w, S_α, and E_ϕ by 0.2287, 0.2103, and 0.3021, respectively, while reducing MAE by 0.2552. Per-image Spearman correlations between $\mathcal{D}_{10}$ and these four improvements are 0.393, 0.629, 0.628, and 0.746. The association with F_β^w also remains positive within CAMO (0.353) and COD10K (0.371). These observations support the conditional interpretation of the spectral analysis: multi-mode aggregation tends to help most when substantial foreground-associated energy lies outside the Fiedler direction. The practical mask is not the oracle-filtered foreground signal in Theorem 1; the agreement concerns the observed dependence of

practical gains on the spectral regime. The experiment therefore provides a connection between the theoretical representation advantage and actual mask quality.

Qualitative evidence across dispersion regimes. Figure 6 illustrates the conditional nature of the observed gains. In the low-dispersion example ($\mathcal{D}_{10} = 0.031$), the Fiedler partition already captures the foreground well, and ten-mode aggregation lowers patch IoU from 0.945 to 0.820. The medium-dispersion example ($\mathcal{D}_{10} = 0.118$) shows a modest improvement from 0.854 to 0.876. In the high-dispersion example ($\mathcal{D}_{10} = 0.927$), the Fiedler partition misses the target, whereas ten-mode aggregation achieves a patch IoU of 0.800. These selected examples illustrate the aggregate trend in Figure 5(b): complementary modes can improve foreground recovery when the Fiedler representation is insufficient, but they need not improve every image. The examples therefore make the conditional advantage concrete, while the population-level evidence is provided by the preceding quantitative analysis.

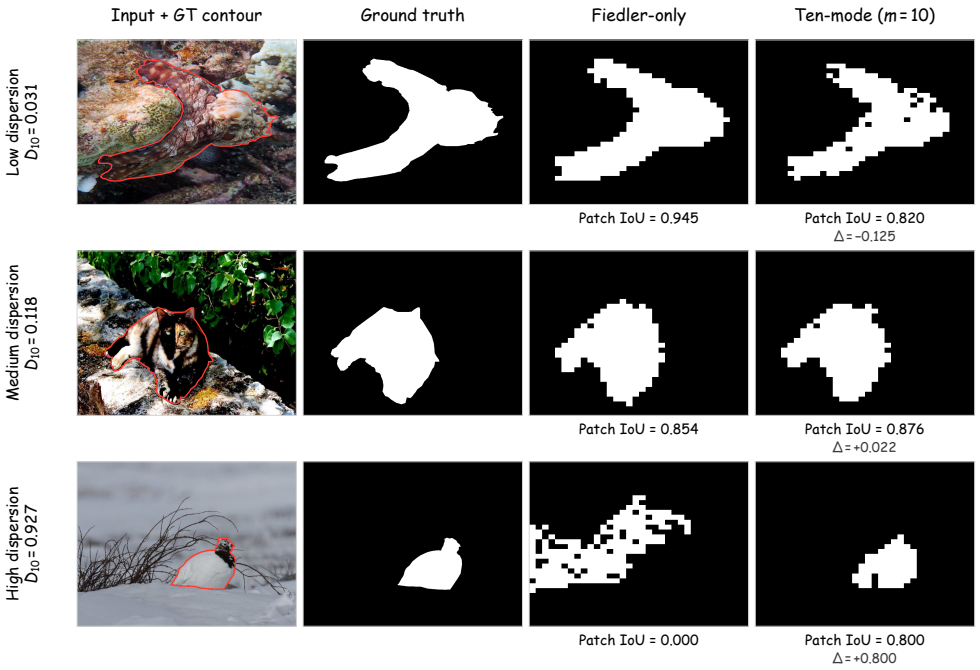


Figure 6: **Qualitative examples across spectral-dispersion regimes.** Rows show individual examples from the low-, medium-, and high-dispersion groups, with $\mathcal{D}_{10} = 0.031$, 0.118, and 0.927, respectively; these are individual values, not group means. Columns show the input with the ground-truth contour, the ground-truth mask, the signed Fiedler partition, and the ten-mode mask. Reported IoUs are measured on the patch grid, with $\Delta = \mathrm{IoU}_{10} - \mathrm{IoU}_{\mathrm{Fiedler}}$; masks are enlarged only for display. The examples illustrate a decrease, a modest gain, and foreground recovery under ten-mode aggregation as dispersion increases. Ground truth is used only for post-hoc analysis and evaluation.

Mode-count saturation and the choice of ten modes. Figure 5(c) evaluates $m \in \{1, 5, 10, 20\}$ using the same non-negative response construction within the sweep. Mean patch-level IoU is 0.4613, 0.5215, 0.5267, and 0.5326, respectively. Here, the $m = 1$ response is proportional to ϕ_2^2 and is distinct from the signed Fiedler partition in the preceding comparison. Increasing the mode count from 10 to 20 raises mean coverage from 0.6588 to 0.7058, but improves patch IoU by only 0.0059. Ten modes already capture 91.7% of the total mean IoU improvement observed between one and twenty modes. Thus, although additional modes retain further foreground energy, their marginal contribution to the resulting mask is much smaller. This behaviour supports $m = 10$ as a compact choice that captures most of the measured benefit. The twenty-mode response remains slightly better on average in this diagnostic sweep; the evidence establishes diminishing returns rather than a universal optimum at ten modes. Taken together, the coverage comparison, dispersion-conditioned gains, qualitative examples, and mode-count sweep provide complementary empirical support for the multi-mode motivation: meaningful foreground energy exists beyond the Fiedler eigenspace, this dispersed energy is associated with practical benefits from aggregation, and most of the measured improvement is obtained using a small set of low-frequency modes.

Algorithm 1 Training and Inference Procedure of DES-UCOD

Require: Unlabelled training set $\mathcal{D} = \{X_n\}_{n=1}^{N}$; frozen encoder $\mathcal{E}$; $m, t, G, \tau_s, \lambda_{\text{ent}}$; training epochs E
Ensure: Trained model parameters θ and prediction map P
Stage I: NEDP Pseudo-Label Generation
Initialise pseudo-label set $\mathcal{P} \leftarrow \emptyset$
for each $X_n \in \mathcal{D}$ **do**
  $\mathbf{Z}_n \leftarrow \mathcal{E}(X_n)$; reshape $\mathbf{Z}_n$ into $\mathbf{F}_n$
  Construct affinity graph and Laplacian using Eq. (10)
  Compute finite-time structural response using Eq. (11)
  Perform structural-to-semantic calibration using Eq. (12)
  Generate pseudo-label $\widetilde{\mathbf{Y}}_n$ using Eq. (13)
  $\mathcal{P} \leftarrow \mathcal{P} \cup \{(X_n, \widetilde{\mathbf{Y}}_n)\}$
end for
Stage II: Model Training
Initialise trainable parameters θ
for epoch $= 1$ to E **do**
  for each mini-batch $\{(X_n, \widetilde{\mathbf{Y}}_n)\} \subset \mathcal{P}$ **do**
    for $g = 1, \ldots, G$ **do**
      Generate aligned view $X_{ng} = T_g(X_n)$ and extract $\mathbf{Z}_{ng} = \mathcal{E}(X_{ng})$
      Compute SC-MoE predictions $\mathbf{P}_{ng}$ using Eqs. (19)–(23)
      Sample $\mathbf{M}_{ng}$ using Eq. (14)
    end for
    Compute cross-view support c_{ngi} using Eq. (15)
    Compute structural reward R_{ng} using Eq. (16)
    Compute relative preference weight ω_{ng} using Eq. (17)
    Compute $\mathcal{L}_{\text{S-RPO}}$ using Eq. (18)
    Compute $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{ent}}$ using Eqs. (24)–(25)
    Update θ by minimising $\mathcal{L}_{\text{total}}$ in Eq. (26)
  end for
end for
Stage III: Inference
Given test image X, compute $\mathbf{Z} \leftarrow \mathcal{E}(X)$
Obtain foreground probability map $\mathbf{P}$ through SC-MoE using Eqs. (19)–(23)
return θ, $\mathbf{P}$

A.3 Methodology

To complement the formulation in Section 3, Algorithm 1 summarizes the complete training and inference procedure of DES-UCOD. During training, NEDP first constructs structural pseudo-labels from the frozen encoder features through non-equilibrium diffusion and structural-to-semantic calibration. The model then processes multiple spatially aligned photometric views with SC-MoE, while S-RPO exploits cross-view support and relative preference weighting to reduce the influence of residual pseudo-label noise. The resulting supervision and preference objectives are jointly optimized while keeping the encoder frozen. During inference, only the frozen encoder and the trained SC-MoE decoder are required to produce the final foreground probability map, without NEDP pseudo-label generation or S-RPO optimization.

A.4 Experiments

A.4.1 Experimental Setup

Implementation details. Our experiments are conducted in PyTorch utilising four NVIDIA RTX 4090 GPUs. Relying on a frozen DINOv2 (Oquab et al., 2023) architecture as the foundation, we optimize the network via Adam with a batch size of 32 and an initial learning rate set to 1×10^{-4}. The entire training phase spans 20 epochs. For consistency, images are standardised to a resolution of 476×476 throughout all training and inference procedures.

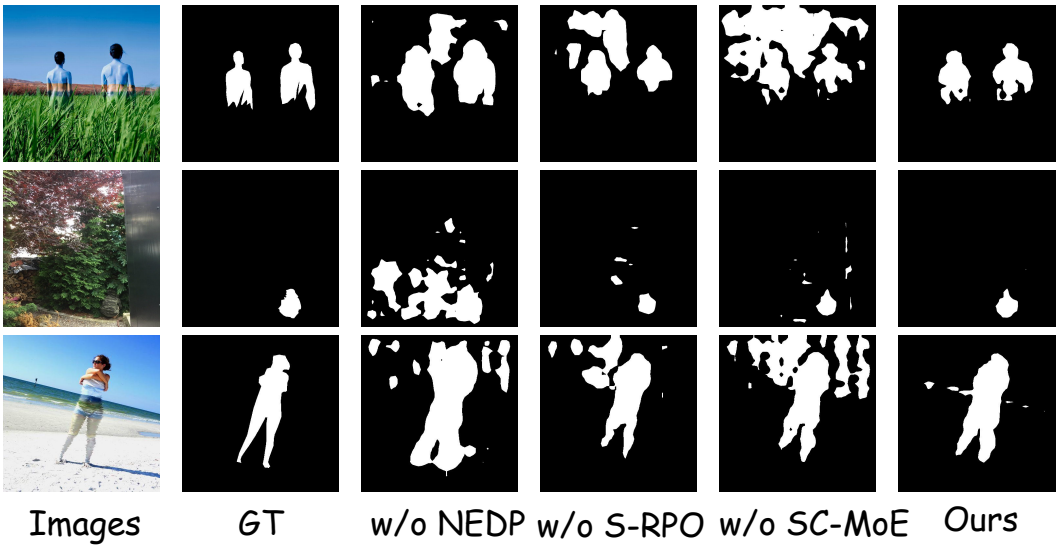


Figure 7: **Qualitative component ablation.** Visual comparison of segmentation results under different component ablations. NEDP facilitates the recovery of more complete object structures, S-RPO improves robustness against noisy pseudo-supervision and suppresses spurious responses, and SC-MoE enhances the modelling of spatially heterogeneous camouflage patterns. By integrating these complementary components, the full model produces more complete foreground regions, cleaner boundaries, and fewer background false positives.

Datasets and Evaluation Metrics. Our training set is composed of 1,000 images from CAMO-Train (Le et al., 2019) and 3,040 images from COD10K (Fan et al., 2020a). DINOv2 (Oquab et al., 2023) is adopted as the visual backbone for producing reliable pseudo labels. During evaluation, we report results on four commonly used camouflage benchmarks, including CHAMELEON with 76 images (Skurowski et al., 2018), CAMO with 250 images, COD10K with 2,026 images, and NC4K with 4,121 images (Lv et al., 2021). Model performance is measured using four representative criteria: Structure Measure (S_α) (Fan et al., 2017), Mean E-measure (E_ϕ) (Fan et al., 2018), Weighted F-measure (F_β^w) (Margolin et al., 2014), and Mean Absolute Error (M) (Perazzi et al., 2012). For pseudo-label quality evaluation, we additionally report adaptive and maximum variants of the F-measure and E-measure to provide a more comprehensive assessment of pseudo-supervision quality.

A.4.2 Further Experimental Analysis

In this section, we provide a more detailed analysis of the experimental results presented in Section 4, including comparisons across different backbones, component-level ablations, efficiency, supervision settings, hyperparameter sensitivity, and pseudo-label quality.

Extended Quantitative Analysis. As shown in Table 1, the DINOv2-based DES-UCOD achieves consistent improvements across all four benchmarks. Compared with DualUCOD on CHAMELEON, our method reduces M from 0.041 to 0.036, while improving F_β^w, E_ϕ, and S_α from 0.754, 0.918, and 0.827 to 0.796, 0.927, and 0.851, respectively. Similar improvements are observed on CAMO, where F_β^w increases from 0.702 to 0.723 while M decreases from 0.083 to 0.081. The improvement is particularly clear on COD10K: compared with DualUCOD, our method reduces M from 0.042 to 0.035 and improves F_β^w, E_ϕ, and S_α by 0.061, 0.031, and 0.032, respectively. On NC4K, DES-UCOD further improves F_β^w from 0.745 to 0.783 and reduces M from 0.052 to 0.047. The consistent improvements across M, F_β^w, E_ϕ, and S_α indicate that the gains are not limited to a single evaluation criterion. In particular, the improvements in F_β^w and S_α suggest better foreground completeness and structural consistency, while the reduction in M reflects lower overall pixel-wise errors. These results are consistent with the complementary roles of NEDP, S-RPO, and SC-MoE in structural discovery, noise-robust learning, and adaptive feature decoding. The backbone comparison provides another useful observation. Although the DINO-based version remains competitive, its performance is lower than the DINOv2-based version on the more challenging benchmarks. For ex-

ample, on COD10K, replacing DINO with DINOv2 improves F_β^w from 0.547 to 0.696 and reduces M from 0.057 to 0.035. On NC4K, F_β^w increases from 0.670 to 0.783. This gap is consistent with the stronger structural representation and feature separability provided by DINOv2, which can lead to more informative graph affinities for NEDP. Nevertheless, the competitive results obtained with DINO indicate that the proposed framework is not restricted to a particular backbone.

Extended qualitative analysis. Fig. 7 visually compares the component ablations. Overall, the full DES-UCOD yields masks closest to the ground truth, featuring complete foregrounds, clean boundaries, and minimal false positives. Specifically, NEDP provides crucial structural pseudo-labels; without it, objects are incompletely recovered and poorly separated. S-RPO enhances robustness against noisy pseudo-supervision, preventing fragmented foregrounds and scattered spurious activations. Meanwhile, SC-MoE adapts to spatially heterogeneous camouflage cues, as removing it increases background interference and produces irregular boundaries in complex scenes. By integrating these components, the complete model reliably recovers full object extents while avoiding noise, perfectly aligning with the quantitative improvements in Tables 2, 3, and 4.

Detailed Component Ablation. Table 2 further reveals different roles of the three proposed components. Removing NEDP causes the largest performance degradation on COD10K, increasing M from 0.035 to 0.052 and decreasing F_β^w from 0.696 to 0.624. Meanwhile, E_ϕ and S_α decrease from 0.887 and 0.793 to 0.814 and 0.752, respectively. This substantial degradation indicates that the quality and completeness of the initial structural pseudo-supervision are particularly important for COD10K. Removing S-RPO produces a different pattern. On COD10K, M increases moderately from 0.035 to 0.040, whereas on NC4K it increases more substantially from 0.047 to 0.058. This suggests that S-RPO becomes particularly useful when predictions contain more locally inconsistent or uncertain regions, because its relative preference mechanism reduces the influence of poorly supported pseudo-label regions. In contrast, removing SC-MoE results in $M = 0.038$ and $F_\beta^w = 0.685$ on COD10K, and $M = 0.052$ and $F_\beta^w = 0.769$ on NC4K. These results show that adaptive spatial decoding provides complementary improvements after reliable pseudo-supervision has been established. Table 3 further isolates the contribution of NEDP. On COD10K, the basic Spectral baseline achieves $M = 0.054$ and $F_\beta^w = 0.622$, while DESPL improves them to 0.040 and 0.641. The complete NEDP further reaches $M = 0.035$ and $F_\beta^w = 0.696$, together with improvements in E_ϕ and S_α. This progressive improvement supports the use of multiple transient low-frequency responses rather than relying on a more limited spectral partition. The effect of SSC is also informative. Removing SSC decreases F_β^w, E_ϕ, and S_α on both datasets. On NC4K, however, M remains unchanged at 0.047. This apparently neutral result on M does not indicate that SSC is ineffective, because the other three metrics still improve with SSC. One possible explanation is that M averages absolute pixel-wise errors and can therefore be less sensitive to improvements in structural organisation when the overall foreground area changes only slightly. In contrast, F_β^w, E_ϕ, and S_α more directly reflect structural and region-level prediction quality. Table 4 confirms that the three experts in SC-MoE provide complementary information. Removing LCE reduces F_β^w from 0.696 to 0.680 on COD10K, indicating the importance of subtle local contrast. Removing RSE further reduces F_β^w to 0.670 and S_α to 0.780, suggesting that regional aggregation is important for spatially separated but coherent object parts. Removing GCE produces $M = 0.039$ and $F_\beta^w = 0.671$ on COD10K and gives the largest degradation in S_α on NC4K among the three expert ablations, showing the benefit of incorporating global context. Interestingly, removing RSE slightly improves M on NC4K from 0.047 to 0.046. However, F_β^w, E_ϕ, and S_α all decrease. This small change in M is therefore not an overall performance gain. Instead, it suggests a metric-specific trade-off: a model can slightly reduce average pixel-wise error while simultaneously producing less complete or less structurally consistent foreground predictions. The full SC-MoE obtains a better balance across all evaluation metrics.

Efficiency and Supervision Analysis. Table 5 shows that DES-UCOD contains 308.97M parameters in total but only 4.60M trainable parameters, as the backbone is frozen during training. Its inference time is 43.98 ms/image, which is close to EASE at 43.72 ms/image and substantially lower than DualUCOD at 85.93 ms/image. DES-UCOD also requires far fewer trainable parameters than DualUCOD, whose trainable parameter count reaches 1063.93M. These comparisons indicate that the proposed modules introduce only a moderate trainable overhead relative to the frozen representation. RISE is substantially faster at 8.61 ms/image and requires fewer FLOPs. However, this efficiency advantage is accompanied by consistently lower segmentation performance than the DINOv2-based

Table 9: Comparison of Pseudo-Label Quality.

Model	*COD10K-Train + CAMO-Train* (4040 images)								
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^m \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$E_\phi^a \uparrow$	$E_\phi^m \uparrow$	$F_\beta^a \uparrow$	$F_\beta \uparrow$
Spectral (Melas-Kyriazi et al., 2022)	0.659	0.708	0.536	0.168	0.522	0.710	0.710	0.537	0.537
DESPL (Liu et al., 2026a)	0.691	0.754	0.561	0.095	0.539	0.756	0.756	0.562	0.562
SDRD (Liu et al., 2026b)	0.598	0.580	0.393	0.100	0.310	0.703	0.711	0.462	0.497
Ours	**0.717**	**0.808**	**0.627**	**0.079**	**0.588**	**0.810**	**0.810**	**0.629**	**0.629**

DES-UCOD. For example, on COD10K, RISE achieves $F_\beta^w = 0.600$ and $M = 0.049$, compared with 0.696 and 0.035 for DES-UCOD. Therefore, the higher computational cost of DES-UCOD is associated with a clear improvement in segmentation quality, resulting in a different accuracy–efficiency operating point rather than a uniform advantage in all efficiency metrics. Table 6 further compares different supervision settings. Despite using no human annotations, DES-UCOD achieves $M = 0.036$ and $E_\phi = 0.927$ on CHAMELEON, outperforming all listed weakly supervised methods on these two metrics. On NC4K, our method similarly achieves the lowest M of 0.047 and the highest E_ϕ of 0.911 among the listed weakly supervised and unsupervised methods. DES-UCOD does not outperform all weakly supervised methods on every metric. For example, on NC4K, its $S_\alpha = 0.827$ is slightly lower than SEE and WS-SAM. Larger gaps are also observed on CAMO and COD10K. These differences are reasonable because weakly supervised approaches can exploit human-provided point, scribble, or other sparse annotations, whereas DES-UCOD relies entirely on automatically discovered pseudo-supervision. The results therefore indicate that DES-UCOD can approach or exceed weakly supervised performance on several metrics without requiring annotation, rather than implying uniform superiority under all evaluation criteria.

Hyperparameter Sensitivity. Table 7 shows that $m = 10$, $t = 10$, $G = 6$, $\tau_s = 0.5$, and $\lambda_{\text{ent}} = 0.1$ provide the strongest overall configuration on COD10K. The effect of m exhibits a clear intermediate optimum. Increasing m from 6 to 10 improves F_β^w from 0.675 to 0.696 and reduces M from 0.039 to 0.035, whereas further increasing m to 12 reduces F_β^w to 0.677. This trend suggests that too few modes may provide insufficient structural coverage, while retaining too many modes can introduce less informative spectral responses. A similar non-monotonic trend appears for the diffusion time t. The selected value $t = 10$ achieves $M = 0.035$ and $F_\beta^w = 0.696$, outperforming both the smaller value $t = 5$ and larger values $t = 15$ and 20. This behaviour is consistent with the non-equilibrium motivation developed in Section 2: a very small diffusion time may provide insufficient spectral filtering, whereas excessive diffusion can overly suppress useful non-stationary responses. For the number of aligned views, increasing G from 2 to 6 improves F_β^w from 0.675 to 0.696, while $G = 8$ decreases it to 0.672. This indicates that additional views are beneficial up to a moderate number, after which redundant or increasingly difficult augmented views may provide diminishing benefits for relative preference estimation. Similarly, $\tau_s = 0.5$ performs better than 0.3, 0.7, and 1.0, suggesting that an intermediate relaxation temperature provides a useful balance between stochastic exploration and stable optimisation. The entropy weight exhibits the same pattern. $\lambda_{\text{ent}} = 0.1$ achieves the best overall results, whereas both smaller and larger values lead to performance degradation. A small entropy weight may provide insufficient regularisation against premature distribution concentration, while an excessively large weight can over-regularise the predictions. Overall, these results show that the selected hyperparameters consistently lie in moderate regimes rather than at extreme values, supporting the intended balance between structural discovery and stochastic exploration.

Comparison of the Quality of Pseudo-Labels. Table 9 presents a direct evaluation of the pseudo-labels generated before subsequent model training. Our method consistently achieves the strongest performance across all reported pseudo-label quality metrics on the combined COD10K-Train and CAMO-Train set. Compared with DESPL, S_α improves from 0.691 to 0.717, E_ϕ from 0.754 to 0.808, and F_β^m from 0.561 to 0.627, while M decreases from 0.095 to 0.079. These correspond to absolute improvements of 0.026, 0.054, and 0.066 in the three similarity metrics, together with a reduction of 0.016 in MAE. The improvement is also consistent across the remaining weighted, adaptive, and maximum F-measure/E-measure variants reported in Table 9. This consistency is important because it indicates that the advantage is not caused by optimising only one particular evaluation criterion. Instead, NEDP produces pseudo-labels with simultaneously improved pixel accuracy, foreground agreement, and structural quality. Compared with the basic Spectral base-

line, the improvement is even more pronounced, especially in M and the F-measure-based metrics. This supports the motivation that a single or limited spectral partition can miss complementary foreground structures, while finite-time multi-mode diffusion recovers richer structural information. The higher-quality pseudo-labels also provide a plausible explanation for the large performance drop observed when NEDP is removed in Table 2: better initial structural supervision provides a stronger foundation for subsequent S-RPO and SC-MoE learning.

A.5 Limitations and Future Work

Despite the strong performance with DINOv2, we observe a noticeable degradation when using DINO, particularly on challenging benchmarks such as CAMO, COD10K, and NC4K. This suggests that the current framework remains sensitive to the structural quality of the pretrained representations used to construct graph affinities and that the effectiveness of non-equilibrium diffusion may vary across feature spaces. Future work will therefore focus on improving backbone robustness by developing representation-adaptive affinity construction and spectral diffusion, exploring multi-layer and multi-scale feature aggregation, and reducing the dependence of pseudo-label quality on a specific pretrained encoder. We also plan to investigate adaptive selection of diffusion parameters from label-free spectral statistics and to evaluate the resulting framework across a broader range of backbones and cross-domain settings.