SEED, DIFFUSE, AND REFINE: RELIABILITY-SCREENED PRIOR LEARNING FOR UNSUPERVISED CAMOUFLAGED OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised Camouflaged Object Detection (UCOD) aims to eliminate the dependence on costly pixel-level annotations, yet remains highly challenging due to the extremely low foreground-background contrast and the absence of reliable supervision. Recent UCOD approaches increasingly exploit priors from foundation models to alleviate annotation scarcity, but often directly propagate these priors without explicitly assessing their reliability, resulting in noisy pseudo supervision and limited robustness in complex camouflage scenarios. In this paper, we revisit UCOD from the perspective of prior reliability and propose a reliability-screened framework that selectively exploits foundation-model knowledge. Instead of directly inheriting model-generated priors, we leverage the rejection-aware capability of SAM3 to perform semantic probing and construct a SAM3-screened seed set for subsequent learning. Based on these screened but limited seeds, we introduce a diffusion-based segmentation head to capture the distributional characteristics of camouflage structures rather than directly fitting deterministic masks. Furthermore, we propose a SAM3-guided boundary refinement strategy, where diffusion predictions are converted into geometric prompts to adapt SAM3 for boundary-aware mask refinement. Finally, we develop a dual-sided distillation scheme that transfers complementary geometric and semantic priors to the final segmentation model. Extensive experiments on three COD benchmarks demonstrate that our method achieves state-of-the-art performance among UCOD methods.

1 INTRODUCTION

Camouflaged Object Detection (COD) aims to identify objects that are visually concealed within complex surroundings through subtle similarities in appearance, texture, and patterns Fan et al. (2021; 2023). Despite its importance in applications such as wildlife monitoring and medical analysis, COD remains challenging due to the extremely low foreground-background contrast, which makes pixel-level annotation expensive and limits the scalability of fully supervised approaches.

Unsupervised Camouflaged Object Detection (UCOD) has emerged as a promising alternative by alleviating the dependence on dense pixel-level annotations. Recent UCOD methods mainly leverage pre-trained foundation models to provide semantic or visual priors for camouflage understanding. Representative approaches exploit MLLMs or large-scale visual models to extract task-relevant semantic cues Chen et al. (2026) and generalized representations Yan et al. (2025), achieving significant progress by leveraging the strong generalization ability of foundation models. Despite their effectiveness, existing methods generally utilize model-derived priors directly, while the reliability of such priors for challenging COD scenarios remains insufficiently explored. Since camouflaged objects exhibit extreme appearance ambiguity even for human observers, responses from foundation models may contain unreliable semantic or geometric priors. Blindly propagating such unreliable information can introduce cascading errors throughout downstream learning.

The closest work in this direction is UCOD-MKD Chen et al. (2026), which emphasizes pseudo-label quality and applies hand-crafted rules (e.g., boundary-touching masks and connected component statistics) to filter SAM-generated masks, as shown in Fig. 1. However, UCOD-MKD still suffers from several limitations: 1) it lacks explicit assessment of the initial semantic priors that guide

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.

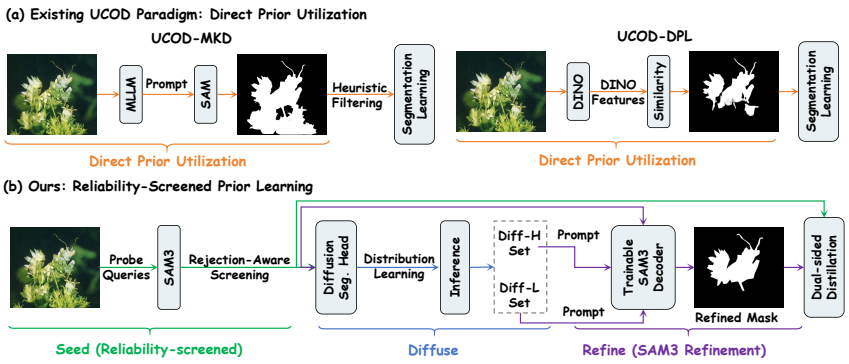

Figure 1: Comparison between existing UCOD paradigms and our reliability-screened Seed-Diffuse-Refine framework. Existing methods, including UCOD-MKD Chen et al. (2026) and UCOD-DPL Yan et al. (2025), directly utilize foundation-model priors without explicit reliability screening. Our framework first screens more reliable priors via SAM3 rejection-aware probing, then learns camouflage structures with diffusion, and finally refines predictions through SAM3-based boundary refinement.

pseudo-mask generation, allowing upstream uncertainty to propagate downstream; 2) its heuristic filtering relies on fixed rules that may lack robustness in complex camouflaged scenes; and 3) discarded samples are abandoned rather than being progressively recovered and exploited.

The key issue in UCOD is not the absence of foundation-model priors, but the lack of explicit prior screening before these priors are propagated through the learning pipeline. Motivated by this observation, we argue that selective and reliability-oriented utilization of model priors is essential for robust UCOD, and propose a new framework built around this principle. Rather than passively inheriting priors from foundation models, we exploit the rejection-aware behavior of SAM3 Carion et al. (2026) as a natural mechanism for reliability-oriented prior screening. Specifically, SAM3 may return no valid mask when a queried concept cannot be effectively grounded in the image. Leveraging this property, we design a set of semantic probe prompts to interrogate SAM3 about potential camouflaged targets. Only samples where SAM3 responds with non-empty grounding masks are retained as **SAM-H**, forming a compact but more reliable seed set for subsequent learning.

Starting from this SAM3-screened seed set, we depart from conventional hard-label segmentation learning and introduce a diffusion-based segmentation head to model the distributional characteristics of camouflaged appearances. This design is motivated by two characteristics of UCOD: the SAM3-screened seed set is necessarily compact due to the difficulty of obtaining reliable priors from generic task-level queries (e.g., "camouflaged object"), and camouflage boundaries are inherently ambiguous due to gradual foreground-background transitions. Directly fitting hard pseudo masks under such conditions may lead to overfitting of noisy boundaries, whereas diffusion-based learning encourages the model to capture robust camouflage structures beyond sparse seed supervision. The trained model is then applied to the full dataset, and its predictions are grouped according to the SAM3 screening status into seed-accepted (Diff-H) and seed-rejected (Diff-L) subsets. Since the diffusion model is trained exclusively on SAM-H, Diff-H inherits the screened prior information from the seed set, while Diff-L provides predictions for samples without explicit SAM3 screening and is further utilized for subsequent refinement.

While the diffusion-based segmentation head effectively captures the distributional characteristics of camouflaged appearances, it may still produce coarse boundaries due to the stochastic nature of diffusion generation. To recover fine-grained spatial details, we introduce a SAM3-guided refinement strategy, where geometric prompts are derived from diffusion predictions and used to adapt SAM3 for boundary-aware mask refinement. This sequential collaboration enables diffusion predictions to be transformed into more accurate geometric priors rather than directly serving as final masks. The refined masks provide full-coverage geometric distillation priors, while semantic logits are selectively retained only from SAM-H samples to preserve reliable semantic guidance. This asymmetric prior utilization strategy expands geometric supervision to the entire dataset while maintaining semantic priors from reliable foundation-model responses. Together, the complementary geometric and semantic priors supervise the final segmentation network through dual-sided distillation.

Our contributions are summarized as follows:

- We identify the reliability issue in existing UCOD prior utilization and introduce a reliability-aware prior screening paradigm that exploits SAM3's rejection-aware behavior to construct a compact but more reliable seed set.
- We introduce a diffusion-based segmentation head tailored for UCOD, which learns robust camouflage structures from SAM3-screened seeds and generates dense predictions beyond the initial sparse supervision.
- We propose a diffusion-guided SAM3 refinement strategy, where diffusion predictions are converted into geometric prompts for boundary-aware SAM3 refinement. The geometric priors complement the selectively retained semantic priors for dual-sided distillation.
- Extensive experiments on COD10K, CAMO, and NC4K demonstrate the effectiveness of our method and establish state-of-the-art performance among UCOD approaches.

2 Related Work

Camouflaged Object Detection. Camouflaged Object Detection (COD) aims to segment objects that are visually concealed within complex surroundings Fan et al. (2020). Fully supervised methods Pang et al. (2022); Luo et al. (2024); He et al. (2025b); Guo et al. (2026) have achieved remarkable progress but rely heavily on costly pixel-level annotations. To reduce annotation dependency, weakly supervised approaches exploit coarse supervision, such as points Chen et al. (2024a), bounding boxes Chen et al. (2024b), and scribbles He et al. (2025a), yet still require manual labeling. Unsupervised Camouflaged Object Detection (UCOD) further removes human annotation by exploiting automatically generated priors. Existing UCOD methods mainly follow two paradigms. *Training-based* approaches typically generate pseudo-labels and optimize segmentation models iteratively. Representative methods include self-supervised adaptation Zhang & Wu (2023), pseudo-label fusion strategies Yan et al. (2025), and teacher-student frameworks with foundation models and mask quality filtering Chen et al. (2026). In contrast, *training-free* approaches Tang et al. (2024); Hu et al. (2024a;b); Du et al. (2025) leverage large-scale foundation models to obtain semantic and spatial prompts for zero-shot segmentation. Despite their differences, both paradigms heavily depend on model-derived priors, while the reliability of such priors in challenging camouflage scenarios remains largely unexplored. Although UCOD-MKD Chen et al. (2026) attempts to improve pseudo-label quality through heuristic filtering, it focuses mainly on downstream mask refinement rather than explicitly screening the initial priors used for pseudo-mask generation. Our work addresses this limitation by introducing a reliability-screened prior utilization paradigm that selectively exploits more reliable foundation-model priors before downstream learning.

Diffusion Models for Segmentation. Diffusion models are generative frameworks that iteratively recover structured outputs from noisy states and have recently been extended beyond image synthesis to dense prediction tasks, including segmentation, edge detection, and depth estimation Ke et al. (2024); Ye et al. (2024). Compared with deterministic decoders, diffusion-based segmentation methods have demonstrated advantages in modeling complex structures and refining predictions through iterative denoising Wu et al. (2024). In the COD domain, FocusDiffuser Zhao et al. (2024a) and CamoDiffusion Sun et al. (2025) explore diffusion formulations for supervised camouflaged object prediction, focusing on boundary refinement and conditional generation. Diffusion models have also been applied to camouflaged image synthesis Luo et al. (2023); Zhao et al. (2024b) to augment training data under challenging visual conditions. Different from existing diffusion-based COD methods that mainly rely on dense supervision, our work introduces a diffusion-based segmentation head into the UCOD setting, where diffusion learning captures robust camouflage structures from SAM3-screened sparse priors, while SAM3-based refinement further improves spatial precision.

Segment Anything Models for COD. SAM Kirillov et al. (2023) is a general-purpose interactive segmentation model that supports various prompt types and has been widely adopted in weakly supervised and unsupervised COD. In weakly supervised scenarios, SAM is typically prompted with sparse annotations, such as points or bounding boxes, to generate pseudo masks He et al. (2025a); Chen et al. (2024b). In unsupervised settings, methods including CVP, GenSAM, ProcMac, and EASE combine MLLMs with other foundation models to derive spatial prompts and employ SAM for segmentation Hu et al. (2024a;b); Du et al. (2025). However, these approaches mainly treat SAM

as a passive segmentation executor, while the reliability of the generated prompts is not explicitly assessed. Recently, SAM3 Carion et al. (2026) extends the SAM family by supporting category-level semantic prompts and introducing a rejection mechanism that enables the model to return no valid segmentation when a queried concept cannot be effectively grounded. In this work, we exploit this rejection-aware behavior as a natural mechanism for prior reliability screening, enabling automatic identification of higher-confidence foundation-model responses without manual annotation.

3 METHOD

3.1 SAM3-SCREENED SEED ACQUISITION

Given an unlabeled image set $\mathcal{D}_u = \{I_i\}_{i=1}^{N}$, we first construct a reliable seed set by exploiting the semantic querying capability of SAM3, as shown in Fig. 2. Unlike existing UCOD methods that directly inherit priors from foundation models, we leverage SAM3's rejection-aware behavior as a natural prior reliability screening mechanism. Specifically, SAM3 may return no valid mask when a queried concept cannot be effectively grounded in the image, avoiding the direct use of unsuccessful responses as pseudo supervision. This property enables automatic identification of more reliable responses without additional human annotation.

Since "camouflaged object" is a highly abstract task-level concept, a single semantic query may not sufficiently retrieve relevant responses from SAM3. Therefore, we construct a query pool $\mathcal{Q}$ containing diverse descriptions of potential camouflaged objects, e.g., $\mathcal{Q} = \{q_1 = \text{"Camouflaged Object"}, q_2 = \text{"Concealed Object"}, \ldots\}$.

These complementary semantic queries provide diverse perspectives for probing potential camouflage targets. For each image, all queries in $\mathcal{Q}$ are independently processed by SAM3. If at least one query produces a valid response, the sample is retained as a candidate seed. When multiple valid responses are obtained, the one with the highest SAM3 response score is selected as the final seed mask M^{seed} and its corresponding semantic embedding Z^{seed}. Formally, for each image I_i, all queries are sequentially fed into SAM3:

$$O_{i,q} = \mathrm{SAM3}(I_i, q), \quad q \in \mathcal{Q}, \tag{1}$$

where $O_{i,q}$ denotes the response of SAM3 for the image-query pair (I_i, q), which can be either a predicted mask $M_{i,q}$ or a rejection response $\emptyset$. Samples receiving valid SAM3 responses are retained as the SAM3-screened seed set:

$$\mathcal{S}_H = \{(I_i, M_i^{seed}, Z_i^{seed}) \mid \exists q \in \mathcal{Q},\ O_{i,q} \neq \emptyset\}. \tag{2}$$

The remaining samples without valid SAM3 responses are collected into $\mathcal{S}_L$. Although $\mathcal{S}_H$ is limited in scale due to the rejection-aware screening process, these SAM3-screened predictions provide a reliable initialization for subsequent camouflage structure learning.

3.2 CAMOUFLAGE DISTRIBUTION LEARNING VIA DIFFUSION

Although the SAM3-screened seed set $\mathcal{S}_H$ provides reliable initialization, its limited scale makes conventional hard-label segmentation learning insufficient for capturing diverse camouflage appearances. This challenge is further amplified in COD, where objects often exhibit gradual foreground-background transitions and ambiguous boundaries. Instead of directly fitting deterministic masks from sparse pseudo labels, we introduce a diffusion-based segmentation head to model the distributional characteristics of camouflage structures. By learning to recover reliable masks from perturbed and noisy structural states, the diffusion process encourages robust structure learning beyond precise boundary fitting.

Feature Extraction and Fusion. We adopt PVTv2 as the visual backbone to extract multi-scale image representations:

$$\{F_1, F_2, F_3, F_4\} = \mathrm{PVTv2}(I). \tag{3}$$

The extracted features are projected into a unified channel dimension and progressively fused to generate the conditional feature:

$$F_c = \mathrm{Fuse}(F_1, F_2, F_3, F_4), \tag{4}$$

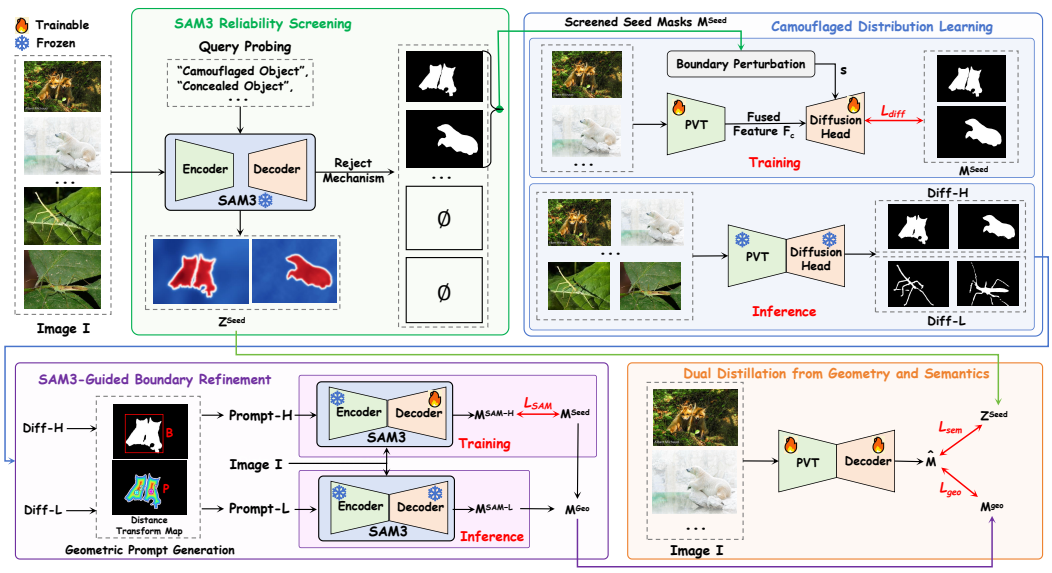

Figure 2: Overall framework of our reliability-screened UCOD method. SAM3 probing selects high-confidence seeds, diffusion learns camouflage structures from screened priors, and SAM3 refinement converts diffusion predictions into geometric priors for dual-sided distillation.

which provides semantic and spatial guidance for the subsequent conditional diffusion process.

Forward Diffusion Process. To alleviate the overfitting problem caused by sparse supervision and ambiguous boundaries, we introduce a boundary perturbation operation $\mathcal{J}(\cdot)$ to construct a boundary-perturbed structural condition from the seed mask:

$$\tilde{M} = \mathcal{J}(M^{seed}), \tag{5}$$

where $\tilde{M}$ represents a perturbed structural prior that preserves the coarse object configuration while introducing controlled variations around object boundaries. Specifically, $\mathcal{J}(\cdot)$ randomly perturbs boundary regions through morphological operations, such as erosion and dilation, to simulate the ambiguous boundaries commonly observed in camouflaged scenes. Unlike treating the perturbed mask as the clean diffusion state, we use $\tilde{M}$ as an auxiliary condition to guide denoising, while the original seed mask remains the reconstruction target. The clean diffusion state is defined as:

$$x_0 = M^{seed}. \tag{6}$$

Following the standard diffusion formulation, Gaussian noise is progressively injected according to a predefined variance schedule $\{\beta_t\}_{t=1}^{T}$. The noisy mask at timestep t is sampled as:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{7}$$

where $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$ controls the retained signal ratio. Equivalently, the noisy state is obtained by:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{8}$$

Conditional Denoising Prediction. The reverse diffusion process reconstructs the clean camouflage mask conditioned on image features and the boundary-perturbed structural condition:

$$p_\theta(x_{t-1}|x_t, \tilde{M}, F_c) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, \tilde{M}, F_c, t), \Sigma_\theta). \tag{9}$$

We adopt the x_0-prediction objective, where the diffusion head estimates the clean seed mask:

$$\hat{M} = D_\theta(x_t, \tilde{M}, F_c, e_t), \tag{10}$$

where e_t denotes the timestep embedding. Based on the predicted clean state $\hat{M}$, the reverse transition mean is computed as:

$$\mu_\theta = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{M} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t. \tag{11}$$

The diffusion head is optimized by reconstructing the SAM3-screened seed masks:

$$\mathcal{L}_{diff} = \mathcal{L}_{structure}(\hat{M}, M^{seed}), \tag{12}$$

where $\mathcal{L}_{structure}$ combines weighted BCE and weighted IoU losses. By introducing $\tilde{M}$ as a boundary-perturbed structural condition, the diffusion model learns robust camouflage structures from noisy observations while reducing overfitting to ambiguous local boundaries.

3.3 SAM3-Guided Boundary Refinement

Although the diffusion-based segmentation head captures the distributional characteristics of camouflaged objects, its predictions may still lack precise boundary localization due to the stochastic nature of diffusion sampling. To address this limitation, we introduce a SAM3-guided refinement strategy, where diffusion predictions are converted into geometric prompts to adapt the SAM3 decoder for boundary-aware mask refinement. Unlike treating SAM as a static pseudo-label generator, our framework exploits the complementary strengths of diffusion-based structure learning and SAM3-based boundary refinement in a sequential manner. Hereafter, we denote a diffusion prediction as M^{Diff}, which corresponds to either **Diff-H** or **Diff-L** according to the SAM3 screening status of the corresponding sample.

Geometric Prompt Generation. For each diffusion prediction M^{Diff}, we derive geometric prompts $\mathcal{G} = \{B, P\}$ consisting of bounding boxes B and foreground points P. The bounding box B is obtained from the minimum enclosing rectangle of the diffusion prediction, providing coarse object localization. For point prompts P, we exploit the distance transform of the diffusion prediction to identify reliable interior regions and sample foreground points with stable spatial cues. This mixed box-point strategy provides SAM3 with complementary localization and interior guidance while reducing sensitivity to ambiguous mask boundaries.

SAM3 Decoder Fine-tuning. To enable SAM3 to adapt to diffusion-generated prompts, we fine-tune its mask decoder while keeping other components fixed. The training pairs are automatically constructed from diffusion predictions generated on $\mathcal{S}_H$:

$$\mathcal{P}_H = \{(M_i^{\text{Diff-H}}, M_i^{seed}) \mid I_i \in \mathcal{S}_H\}. \tag{13}$$

where $M_i^{\text{Diff-H}}$ serves as the geometric prompt source and M_i^{seed} provides the SAM3-screened supervision target. The decoder is optimized with:

$$\mathcal{L}_{SAM}{=}\mathcal{L}_{BCE}(M^{\text{SAM-H}}, M^{seed}) + \mathcal{L}_{Dice}(M^{\text{SAM-H}}, M^{seed}), \tag{14}$$

where $M^{\text{SAM-H}}$ denotes the mask generated by SAM3 with diffusion-derived geometric prompts. By learning from SAM3-screened seed samples, the decoder learns to transform coarse diffusion structures into boundary-refined masks.

SAM3 Refinement on Unscreened Samples. After decoder optimization, the fine-tuned SAM3 is applied to samples in $\mathcal{S}_L$. For each diffusion prediction $M^{\text{Diff-L}}$, geometric prompts are generated using the same strategy:

$$M^{\text{SAM-L}} = \text{SAM3}(I, \mathcal{G}(M^{\text{Diff-L}})). \tag{15}$$

The refined masks from $\mathcal{S}_L$ are combined with the SAM3-screened seed masks from $\mathcal{S}_H$ to form a full-coverage geometric distillation prior:

$$M^{geo} = \{M_i^{seed} | I_i \in \mathcal{S}_H\} \cup \{M_i^{\text{SAM-L}} | I_i \in \mathcal{S}_L\}. \tag{16}$$

Instead of forcing semantic responses from SAM3-rejected samples, we retain semantic logits only from SAM3-screened samples:

$$Z^{sem} = Z^{seed}, \quad I \in \mathcal{S}_H. \tag{17}$$

This asymmetric prior utilization strategy expands geometric supervision to the full dataset while preserving semantic priors only from reliable SAM3-screened responses. Together, the complementary geometric and semantic priors supervise the final network through dual-sided distillation.

Table 1: Quantitative comparison with representative COD methods on CAMO, COD10K, and NC4K. "F", "S", "B", "P", and "U" denote fully-supervised, scribble-supervised, box-supervised, point-supervised, and unsupervised methods, respectively. **Bold** and underline indicate the best and second-best results within each supervision group.

Methods	Sup.	CAMO				COD10K				NC4K			
		MAE↓	S_m↑	E_m↑	F_β^w↑	MAE↓	S_m↑	E_m↑	F_β^w↑	MAE↓	S_m↑	E_m↑	F_β^w↑
Fully-Supervised Methods													
ZoomNet Pang et al. (2022) CVPR'22	F	0.066	0.820	0.892	0.752	0.029	0.838	0.911	0.729	0.043	0.853	0.896	0.784
CamoFocus Khan et al. (2024) WACV'24	F	**0.044**	0.870	0.924	0.842	0.022	0.868	0.931	0.802	0.031	0.886	0.932	0.853
RUN He et al. (2025b) ICML'25	F	0.045	**0.877**	0.934	**0.853**	**0.021**	0.878	0.941	0.808	**0.030**	**0.892**	0.940	**0.866**
SAM-DSA Liu et al. (2025) ICCV'25	F	0.047	0.866	**0.948**	0.839	0.023	**0.881**	**0.969**	**0.817**	0.032	0.889	**0.963**	0.857
Weakly-supervised Methods													
SSE He et al. (2025a) TPAMI'25	S	0.090	0.765	0.826	0.742	0.036	0.807	0.883	0.724	0.051	0.836	0.891	0.805
SAM-COD Chen et al. (2024b) ECCV'24	S	**0.060**	0.836	**0.903**	0.779	0.029	0.833	0.904	0.728	0.039	0.859	0.912	0.795
PCOD Chen et al. (2024a) ECCV'24	P	0.074	0.798	0.872	0.721	0.042	0.784	0.859	0.650	0.051	0.822	0.889	0.748
SAM-COD Chen et al. (2024b) ECCV'24	B	0.062	**0.837**	0.901	**0.786**	**0.028**	**0.842**	**0.914**	**0.745**	**0.037**	**0.867**	**0.923**	**0.813**
Unsupervised Methods													
UCOS-DA Zhang & Wu (2023) ICCVW'23	U	0.129	0.685	0.782	0.584	0.085	0.670	0.751	0.482	0.084	0.741	0.824	0.637
UCOD-DPL Yan et al. (2025) CVPR'25	U	0.108	0.706	0.801	0.621	0.059	0.727	0.822	0.577	0.074	0.761	0.851	0.680
DualUCOD Liu et al. (2026) ICML'26	U	0.099	0.723	0.820	0.629	0.061	0.704	0.795	0.535	0.068	0.764	0.852	0.671
ProMaC Hu et al. (2024b) NIPS'24	U	0.098	0.775	0.832	0.698	0.045	0.795	0.861	0.624	0.073	0.779	0.845	0.707
EASE Du et al. (2025) CVPR'25	U	0.166	0.653	0.737	0.563	0.109	0.673	0.732	0.514	0.108	0.728	0.790	0.633
UCOD-MKD Chen et al. (2026) CVPR'26	U	**0.071**	0.809	**0.875**	**0.753**	0.031	0.835	0.908	0.740	0.040	0.857	0.918	0.803
Ours	U	0.078	**0.815**	0.866	0.744	**0.030**	**0.843**	**0.914**	**0.744**	**0.039**	**0.865**	**0.921**	**0.811**

3.4 Dual Distillation from Geometry and Semantics

We train a final segmentation network comprising a PVTv2-B4 encoder and a lightweight decoder through dual-sided distillation from full-coverage geometric priors and SAM3-screened semantic priors. Unlike conventional pseudo-label learning with a single hard supervision source, this strategy combines complementary spatial and semantic knowledge.

Geometric Distillation. The geometric prior M^{geo} provides mask-level supervision for all unlabeled samples:

$$\mathcal{L}_{geo} = \mathcal{L}_{structure}(\hat{M}, M^{geo}), \tag{18}$$

where $\hat{M}$ denotes the prediction of the final segmentation network. Benefiting from SAM3-guided refinement, M^{geo} provides full-coverage spatial localization supervision, including SAM3-screened seed masks for $\mathcal{S}_H$ samples and SAM3-refined masks for $\mathcal{S}_L$ samples.

Semantic Distillation. Although semantic information provides high-level guidance beyond pixel-level localization, unreliable semantic responses may introduce extra noise. Following our principle of selective prior utilization, semantic distillation is only applied to SAM3-screened samples:

$$\mathcal{L}_{sem} = \mathcal{L}_{MSE}(\hat{M}, Z^{seed}) + \mathcal{L}_{L1}(\hat{M}, Z^{seed}), \quad I_i \in \mathcal{S}_H, \tag{19}$$

where Z^{seed} denotes the semantic logits retained from SAM3-screened responses.

The final optimization objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{geo} + \mathcal{L}_{sem}. \tag{20}$$

The geometric branch expands spatial supervision to the entire dataset, while the semantic branch provides category-level guidance only from reliable SAM3-screened responses. This asymmetric design prevents unreliable semantic information from being propagated while maximizing the utilization of reliable foundation-model priors.

4 Experiments

4.1 Experimental Settings

Training and Test Sets. Following previous UCOD methods, we use the training splits of CAMO and COD10K as the unlabeled training set $\mathcal{D}_u$. We evaluate our method on three widely used COD benchmarks: CAMO Le et al. (2019) with 250 test images, COD10K Fan et al. (2020) with 2026 test images, and NC4K Lv et al. (2021) with 4121 test images.

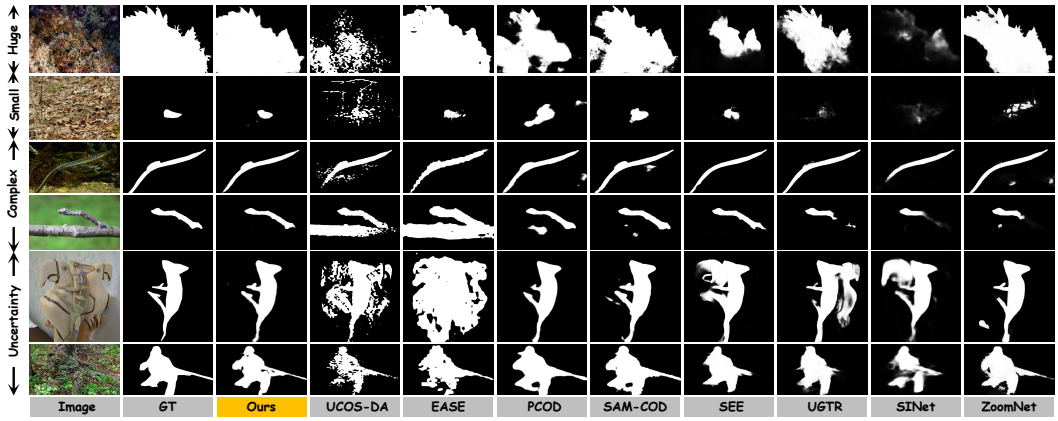

Figure 3: **Visual comparison of our method with existing approaches under challenging camouflage scenarios.** Our method produces more complete object masks with sharper boundaries and preserves fine-grained details of camouflaged targets.

Table 2: Ablation study of individual components on COD10K. Screening, Diffusion, Refine, and Distill denote seed screening, diffusion learning, boundary refinement, and dual-sided distillation, respectively.

Settings				Metrics			
Screening	Diffusion	Refine	Distill	MAE↓	S_m↑	E_m↑	F_β^w↑
-	-	-	-	0.085	0.731	0.807	0.664
✓	-	-	-	0.041	0.790	0.864	0.706
✓	✓	-	-	0.036	0.809	0.882	0.719
✓	✓	✓	-	0.032	0.833	0.905	0.738
✓	✓	✓	✓	**0.030**	**0.843**	**0.914**	**0.744**

Table 3: Semi-supervised comparison with 5% and 10% labeled data. Our variant uses ground-truth seeds, SAM refinement, and SAM logits for soft distillation.

Method	Labeled	MAE↓	S_m↑	E_m↑	F_β^w↑
CamoTeacher Lai et al. (2024)	5%	0.050	0.745	0.827	0.583
ST-SAM Hu et al. (2025)	5%	0.031	0.834	0.871	0.709
Ours	5%	**0.028**	**0.846**	**0.885**	**0.720**
CamoTeacher Lai et al. (2024)	10%	0.049	0.759	0.836	0.594
ST-SAM Hu et al. (2025)	10%	0.029	0.837	0.882	0.723
Ours	10%	**0.027**	**0.849**	**0.893**	**0.731**

Evaluation Protocol. Following Fan et al. (2020), we adopt S-measure (S_m) Fan et al. (2017), mean E-measure (E_m) Fan et al. (2018), weighted F-measure (F_β^ω) Margolin et al. (2014), and mean absolute error (MAE) as evaluation metrics.

Implementation Details. We implement our method in PyTorch on a single NVIDIA RTX A6000 using the official SAM3 checkpoint and pretrained PVTv2-B4 weights. Only the SAM3 mask decoder is fine-tuned; its image and prompt encoders remain frozen. Diffusion training uses 512×512 inputs, 50 epochs, batch size 16, and an initial learning rate of 10^{-3} with cosine decay. SAM3 refinement and final distillation use 20 and 100 epochs with batch sizes 4 and 16, respectively, both with a learning rate and weight decay of 10^{-4}. Diffusion inference uses $T = 10$ sampling steps.

4.2 Main Results

Quantitative Analysis. Tab. 1 compares our method with recent COD approaches under different supervision settings. Our method achieves state-of-the-art UCOD performance on COD10K and NC4K, outperforming UCOD-MKD by 0.008 in S_m on both benchmarks. It achieves the best UCOD results on 9 of 12 metrics across the three benchmarks while remaining competitive on CAMO, supporting the effectiveness of selective prior utilization.

Qualitative Analysis. Fig. 3 compares our method with representative UCOD and WSCOD approaches across challenging object scales, backgrounds, and boundaries. Our method produces more complete regions and finer contours, benefiting from diffusion-based structure learning and SAM3-guided refinement.

Analysis of SAM3-based Seed Screening. We compare raw SAM3 predictions on accepted and rejected training samples, using ground-truth masks only for evaluation. As shown in Tab. 5, accepted samples achieve better quality across all metrics, with MAE of 0.017 versus 0.040 for rejected samples, supporting the effectiveness of screening in selecting higher-quality supervision.

Table 4: Ablation study of different query and screening strategies for seed acquisition.

Query Strategy	Screening	MAE↓	S_m↑	E_m↑	F_β^w↑
Single Query	-	0.085	0.731	0.807	0.664
Query Pool	-	0.074	0.748	0.822	0.677
Query Pool	✓	**0.041**	**0.790**	**0.864**	**0.706**

Table 5: Reliability analysis of SAM3-based seed screening.

Group	#Samples	Ratio	MAE↓	S_m↑	E_m↑	F_β^w↑
All samples	4040	100%	0.032	0.870	0.922	0.830
SAM3-Accepted	1591	39.4%	0.017	0.925	0.969	0.922
SAM3-Rejected	2449	60.6%	0.040	0.840	0.893	0.786

Table 6: Ablation study of the diffusion-based structure learning and boundary perturbation strategy.

Seg. Head	Boundary Perturb.	MAE↓	S_m↑	E_m↑	F_β^w↑
Baseline Decoder	-	0.039	0.796	0.866	0.708
Diffusion Head	-	0.037	0.805	0.877	0.716
Diffusion Head	✓	**0.036**	**0.809**	**0.882**	**0.719**

Table 7: Effect of Geometry-Semantic Distillation.

Geo Distill	Sem Distill	MAE↓	S_m↑	E_m↑	F_β^w↑
✓	-	0.032	0.833	0.905	0.738
-	✓	0.035	0.811	0.884	0.720
✓	✓	**0.030**	**0.843**	**0.914**	**0.744**

4.3 Ablation Study

Ablation Settings. We conduct ablations on COD10K to evaluate seed screening, diffusion learning, SAM3 refinement, and dual-sided distillation, along with query probing, boundary perturbation, and diffusion sampling steps.

Effectiveness of Each Component. Tab. 2 summarizes component contributions. The baseline binarizes SAM3 semantic logits from the "Camouflaged Object" query into pseudo masks to train a conventional segmentation model. Seed screening improves performance by selecting higher-quality responses, while diffusion learns robust camouflage structures from screened seeds. SAM3-guided refinement improves boundary localization, and dual-sided distillation achieves the best performance by combining geometric and semantic priors.

Generalization Beyond SAM3. To evaluate our framework beyond SAM3, we replace SAM3 with SAM and use ground-truth masks from 5% or 10% of training samples as reliable seeds. We retain diffusion-based structure learning, geometric-prompted refinement, and dual-sided distillation, using SAM logits as soft targets. As shown in Tab. 3, this variant outperforms CamoTeacher and ST-SAM across all four metrics at both labeling ratios, supporting its effectiveness beyond SAM3-based seed acquisition and refinement.

Effectiveness of SAM3-based Seed Screening. Tab. 4 compares different seed acquisition strategies. The query pool improves performance by providing broader semantic coverage, while SAM3 rejection-aware screening further improves results by filtering less reliable responses. These results demonstrate the importance of constructing higher-quality seeds before subsequent learning.

Effect of Diffusion-based Structure Learning. Tab. 6 compares a standard decoder with the diffusion head under the same SAM3 refinement and final distillation pipeline, varying only the segmentation head and boundary perturbation. Diffusion improves performance by learning robust camouflage structures from screened seeds, while boundary perturbation further reduces reliance on ambiguous boundary details.

Effect of Geometry-Semantic Distillation. We further analyze geometric and semantic distillation in Tab. 7. Combining both achieves the best performance, validating their complementary roles.

5 Conclusion

In this work, we revisit UCOD from the perspective of prior reliability and reveal that directly propagating foundation-model knowledge may introduce noisy supervision. To address this issue, we propose a reliability-screened UCOD framework that selectively exploits foundation-model priors through a Seed-Diffuse-Refine paradigm. Our framework first constructs SAM3-screened seed priors via rejection-aware semantic probing, then learns robust camouflage structures through diffusion-based segmentation, and further improves boundary localization through SAM3-guided refinement and dual geometry-semantic distillation. Extensive experiments on CAMO, COD10K, and NC4K demonstrate the superiority of our method over existing UCOD approaches.

AI USE STATEMENT

We used ChatGPT to assist with manuscript wording, language polishing, text compression, and LaTeX table formatting. All AI-assisted revisions were reviewed and edited by the authors to ensure accuracy and preserve the intended technical meaning. The authors take full responsibility for the final content of this paper.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *ICLR*, pp. 138846–138923, 2026.

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

Huafeng Chen, Pengxu Wei, Guangqian Guo, and Shan Gao. Sam-cod: Sam-guided unified framework for weakly-supervised camouflaged object detection. In *ECCV*, pp. 315–331. Springer, 2024b.

Huafeng Chen, Chenguang Zhu, Yueming Lyu, and Caifeng Shan. Beyond weak supervision: Mllms-guided graded knowledge distillation for unsupervised camouflaged object detection. In *CVPR*, pp. 27547–27557, 2026.

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 *CVPR*, pp. 19271–19282, 2025.

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

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, pp. 698–704, 2018.

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

Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE TPAMI*, 44(10):6024–6042, 2021.

Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *Visual Intelligence*, 1(1):16, 2023.

Guangqian Guo, Pengfei Chen, Yong Guo, Huafeng Chen, Boqiang Zhang, and Shan Gao. Boosting segment anything model to generalize visually non-salient scenarios. *IEEE TIP*, 2026.

Chunming He, Kai Li, Yachao Zhang, Ziyun Yang, Youwei Pang, Longxiang Tang, Chengyu Fang, Yulun Zhang, Linghe Kong, Xiu Li, et al. Segment concealed objects with incomplete supervision. *IEEE TPAMI*, 2025a.

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 *International Conference on Machine Learning*, pp. 22853–22864. PMLR, 2025b.

Jian Hu, Jiayi Lin, Shaogang Gong, and Weitong Cai. Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In *AAAI*, volume 38, pp. 12511–12518, 2024a.

Jian Hu, Jiayi Lin, Junchi Yan, and Shaogang Gong. Leveraging hallucinations to reduce manual prompt dependency in promptable segmentation. In *NeurIPS*, pp. 107171–107197, 2024b.

Xihang Hu, Fuming Sun, Jiazhe Liu, Feilong Xu, and Xiaoli Zhang. St-sam: Sam-driven self-training framework for semi-supervised camouflaged object detection. In *ACMMM*, pp. 8194–8203, 2025.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, pp. 9492–9502, 2024.

Abbas Khan, Mustaqeem Khan, Wail Gueaieb, Abdulmotaleb El Saddik, Giulia De Masi, and Fakhri Karray. Camofocus: Enhancing camouflage object detection with split-feature focal modulation and context refinement. In *WACV*, pp. 1434–1443, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pp. 4015–4026, 2023.

Xunfa Lai, Zhiyu Yang, Jie Hu, Shengchuan Zhang, Liujuan Cao, Guannan Jiang, Zhiyu Wang, Songan Zhang, and Rongrong Ji. Camoteacher: Dual-rotation consistency learning for semi-supervised camouflaged object detection. In *ECCV*, pp. 438–455, 2024.

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.

Jiaming Liu, Linghe Kong, and Guihai Chen. Improving sam for camouflaged object detection via dual stream adapters. In *ICCV*, pp. 21906–21916, 2025.

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 *ICML*. PMLR, 2026.

Xue-Jing Luo, Shuo Wang, Zongwei Wu, Christos Sakaridis, Yun Cheng, Deng-Ping Fan, and Luc Van Gool. Camdiff: Camouflage image augmentation via diffusion. *CAAI Artificial Intelligence Research*, 2023.

Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. Vscode: General visual salient and camouflaged object detection with 2d prompt learning. In *CVPR*, pp. 17169–17180, 2024.

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 *CVPR*, pp. 11591–11601, 2021.

Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, pp. 248–255, 2014.

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

Ke Sun, Zhongxi Chen, Xianming Lin, Xiaoshuai Sun, Hong Liu, and Rongrong Ji. Conditional diffusion models for camouflaged and salient object detection. *IEEE TPAMI*, 47(4):2833–2848, 2025.

Lv Tang, Peng-Tao Jiang, Zhi-Hao Shen, Hao Zhang, Jin-Wei Chen, and Bo Li. Chain of visual perception: Harnessing multimodal large language models for zero-shot camouflaged object detection. In *ACMMM*, pp. 8805–8814, 2024.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational visual media*, 8(3):415–424, 2022.

Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical imaging with deep learning*, pp. 1623–1639. PMLR, 2024.

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 *CVPR*, pp. 30365–30375, 2025.

Yunfan Ye, Kai Xu, Yuhang Huang, Renjiao Yi, and Zhiping Cai. Diffusionedge: Diffusion probabilistic model for crisp edge detection. In *AAAI*, volume 38, pp. 6675–6683, 2024.

Yi Zhang and Chengyi Wu. Unsupervised camouflaged object segmentation as domain adaptation. In *ICCV*, pp. 4334–4344, 2023.

Jianwei Zhao, Xin Li, Fan Yang, Qiang Zhai, Ao Luo, Zicheng Jiao, and Hong Cheng. Focusdiffuser: Perceiving local disparities for camouflaged object detection. In *ECCV*, pp. 181–198. Springer, 2024a.

Pancheng Zhao, Peng Xu, Pengda Qin, Deng-Ping Fan, Zhicheng Zhang, Guoli Jia, Bowen Zhou, and Jufeng Yang. Lake-red: Camouflaged images generation by latent background knowledge retrieval-augmented diffusion. In *CVPR*, pp. 4092–4101, 2024b.

APPENDIX

Contents

A. SAM3 Seed Screening Details 13
A.1 Query Pool 13
A.2 Query Selection Strategy 13
A.3 SAM Screening Statistics 14
B. Diffusion Head Details 14
B.1 Architecture 14
B.2 Feature Fusion Details 15
B.3 Noise Schedule and Sampling Details 15
C. Boundary Perturbation Details 15
D. SAM3 Decoder Fine-tuning Details 16
D.1 Prompt Generation 16
D.2 SAM3 Decoder Training Details 16
E. More Ablation Studies 16
E.1 Seed Ratio Analysis 16
E.2 Diffusion Step Ablation 16
E.3 Boundary Perturbation Ablation 17
F. Parameter Comparison with Conventional Decoder 17

A SAM3 SEED SCREENING DETAILS

A.1 QUERY POOL

As described in the main paper, a single semantic query may not sufficiently activate SAM3's recognition capability for the abstract concept of camouflaged objects. Therefore, we construct a query pool containing diverse semantic descriptions related to camouflage, concealment, and hidden objects. These queries provide complementary semantic perspectives and are independently fed into SAM3 for each image. If multiple queries produce valid responses, the response with the highest SAM3 score is selected as the final seed annotation. The complete query pool used in all experiments is listed in Tab. A1. The same query pool is applied to all datasets without image-specific modification.

A.2 QUERY SELECTION STRATEGY

For each image, all queries in the query pool are independently fed into SAM3 to obtain potential camouflage responses. When multiple queries produce valid masks, we select the response with the highest SAM3 score as the final seed annotation. This strategy allows different semantic descriptions to complement each other while avoiding the bias introduced by manually selecting a specific query. It is worth noting that multiple queries do not require repeated image encoding in SAM3. Since the image embedding can be computed once and reused for different queries, additional queries only involve the corresponding text encoding and mask decoding processes. Therefore, the computational overhead introduced by the query pool is relatively limited compared with repeatedly processing the entire image.

Table A1: The complete semantic query pool used for SAM3-based seed screening.

ID	Query
q1	”Camouflaged Object”
q2	”Concealed Object”
q3	”Hidden Object”
q4	”Camouflaged Animal”
q5	”Concealed Animal”
q6	”Hidden Animal”
q7	”Animal”

Table A2: Statistics and segmentation quality comparison between SAM3-accepted and SAM3-rejected samples on the training splits of CAMO and COD10K.

Group	#Samples	Ratio	MAE↓	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$
CAMO (1000)						
All samples	1000	100%	0.042	0.864	0.909	0.834
SAM3-Accepted	536	53.6%	0.023	0.920	0.963	0.916
SAM3-Rejected	464	46.4%	0.063	0.800	0.846	0.739
COD10K (3040)						
All samples	3040	100%	0.028	0.871	0.923	0.828
SAM3-Accepted	1055	34.7%	0.016	0.912	0.960	0.888
SAM3-Rejected	1985	65.3%	0.035	0.850	0.904	0.796

A.3 SAM Screening Statistics

To further analyze the effect of SAM3-based screening, we report the detailed statistics of SAM3-accepted and SAM3-rejected samples on the training splits of CAMO and COD10K. As shown in Tab. A2, SAM3-accepted samples consistently achieve better segmentation quality than SAM3-rejected samples across both datasets. For example, on CAMO, SAM3-accepted samples achieve an MAE of 0.023, compared with 0.063 for rejected samples. Similarly, on COD10K, the accepted subset obtains an MAE of 0.016, outperforming the rejected subset with an MAE of 0.035. These results further demonstrate that SAM3 rejection-aware screening can effectively select higher-quality foundation-model responses for subsequent learning.

B Diffusion Head Details

B.1 Architecture

The diffusion-based segmentation head takes the noisy mask x_t, timestep embedding e_t, and image condition feature F_c as inputs to predict the clean mask state x_0. The image condition is extracted by the PVTv2-B4 backbone and contains four-scale features:

$$\{F_1, F_2, F_3, F_4\} = \mathrm{PVTv2}(I). \tag{21}$$

Each feature map is first projected into 256 channels using a 1×1 convolution. To aggregate multi-scale spatial information, F_2, F_3, and F_4 are bilinearly resized to the resolution of F_1 and progressively fused through concatenation followed by a 1×1 convolution, producing the conditional feature F_c. For the noisy mask input, we employ a lightweight mask encoder consisting of a 7×7 convolution with stride 4, followed by time-conditioned residual blocks and a 3×3 convolution. The timestep information is encoded using sinusoidal positional embeddings followed by a two-layer MLP. The encoded mask feature and image condition feature are concatenated and processed by two upsampling decoder blocks to recover the clean mask prediction:

$$\hat{M} = D_\theta(x_t, \tilde{M}, F_c, e_t). \tag{22}$$

The diffusion head adopts an x_0-prediction objective and is optimized using the same structure loss as described in the main paper, which combines weighted BCE loss and weighted IoU loss.

B.2 Feature Fusion Details

The multi-scale features extracted by PVTv2-B4 Wang et al. (2022) are fused to provide image conditions for the diffusion head. Specifically, each feature map is first projected into a unified channel dimension:

$$\bar{F}_i = \phi_i(F_i), \quad i \in \{1, 2, 3, 4\}, \tag{23}$$

where $\phi_i(\cdot)$ denotes a 1×1 convolution and all projected features contain 256 channels. Since different stages have different spatial resolutions, $\bar{F}_2$, $\bar{F}_3$, and $\bar{F}_4$ are resized to the resolution of $\bar{F}_1$ via bilinear interpolation. The aligned features are then concatenated and compressed by another 1×1 convolution:

$$F_c = \psi(\mathrm{Concat}[\bar{F}_1, \mathrm{Up}(\bar{F}_2), \mathrm{Up}(\bar{F}_3), \mathrm{Up}(\bar{F}_4)]), \tag{24}$$

where $\psi(\cdot)$ denotes the channel compression operation. The resulting feature F_c serves as the image condition for the conditional diffusion process.

B.3 Noise Schedule and Sampling Details

Following the standard diffusion formulation, we adopt a discrete variance schedule $\{\beta_t\}_{t=1}^{T}$ to gradually inject Gaussian noise into the clean mask. Specifically, the noise variance is linearly increased from β_{start} to β_{end}:

$$\beta_t = \beta_{start} + \frac{t-1}{T-1}(\beta_{end} - \beta_{start}), \quad t = 1, \ldots, T, \tag{25}$$

where $\beta_{start} = 1 \times 10^{-4}$ and $\beta_{end} = 0.02$. The corresponding noise retention coefficient is defined as:

$$\alpha_t = 1 - \beta_t, \tag{26}$$

and

$$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i. \tag{27}$$

During training, the timestep t is randomly sampled from $\{1, \ldots, T\}$, and the noisy mask x_t is generated according to the forward diffusion process. During inference, we perform the reverse denoising process with $T = 10$ sampling steps to recover the predicted clean mask. The diffusion head operates on 512×512 inputs, consistent with the main experimental setting.

C Boundary Perturbation Details

The boundary perturbation function $\mathcal{J}(\cdot)$ is designed to simulate the boundary ambiguity commonly observed in camouflaged objects. Given a SAM3-screened seed mask M^{seed}, we construct the perturbed structural condition $\tilde{M}$ through two complementary perturbation stages: contour-level perturbation and morphology-level perturbation.

In the contour-level stage, object contours are first extracted using OpenCV. For each valid contour, a consecutive contour segment is randomly removed, and a subset of contour vertices is sampled to reconstruct a coarse object shape. The intermediate contour mask is then further perturbed through local morphological operations. Specifically, random rectangular regions are selected from the mask, and dilation or erosion is randomly applied with equal probability. The structuring element is randomly chosen from rectangular and elliptical kernels, with kernel sizes sampled from $[3, 9]$.

The perturbation process is iteratively performed until the IoU between the perturbed mask and the intermediate contour mask reaches the randomly sampled target range. In our experiments, the target IoU is uniformly sampled from $[0.8, 1.0]$. We perform at most 250 perturbation rounds, with four randomly selected local regions processed in each round. The final perturbed mask $\tilde{M}$ preserves the coarse object configuration while introducing controlled boundary uncertainty for diffusion learning.

Table A3: Effect of different SAM3-screened seed ratios on COD10K.

Seed Ratio	MAE↓	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$
50%	0.035	0.817	0.893	0.725
75%	0.032	0.831	0.903	0.738
100%	0.030	0.843	0.914	0.744

D SAM3 Decoder Fine-tuning Details

D.1 Prompt Generation

For SAM3 decoder refinement, we generate geometric prompts from diffusion predictions using a combination of bounding boxes and foreground points. The bounding box is obtained by computing the tight enclosing rectangle of all foreground pixels in the diffusion mask. For point prompts, we aim to select points from stable interior regions rather than ambiguous boundaries. Specifically, the diffusion prediction is first binarized with a fixed threshold, followed by a distance transform over the binary mask. Pixels with larger distance transform values are farther from object boundaries and are therefore selected as candidate foreground points. For samples containing a single connected component, we select three foreground points. The first point is chosen as the pixel with the maximum distance transform value, while the remaining points are selected using farthest point sampling to encourage spatial diversity. For samples containing multiple connected components, one point is selected from each component according to the maximum distance transform value within the corresponding region. The generated box-point prompts provide complementary coarse localization and interior guidance for SAM3 refinement, while reducing the influence of uncertain boundary regions.

D.2 SAM3 Decoder Training Details

During SAM3 refinement, only the mask decoder is fine-tuned, while the image encoder and prompt encoder remain frozen. The training samples are constructed from the SAM3-screened seed set $\mathcal{S}_H$, where diffusion predictions provide geometric prompts and the original SAM3 seed masks serve as supervision targets. The decoder is optimized using a combination of binary cross-entropy loss and Dice loss:

$$\mathcal{L}_{SAM} = \mathcal{L}_{BCE} + \mathcal{L}_{Dice}. \tag{28}$$

The same fine-tuned decoder is then applied to SAM3-rejected samples using diffusion-derived geometric prompts for mask refinement.

E More Ablation Studies

E.1 Seed Ratio Analysis

To investigate the robustness of our diffusion-based learning under limited seed supervision, we evaluate the effect of different seed ratios on COD10K. Specifically, we randomly sample different proportions of the SAM3-screened seed set $\mathcal{S}_H$ while keeping the overall training pipeline unchanged. The sampled seeds are used for diffusion training, and the remaining unlabeled samples are recovered through the same refinement and distillation process.

As shown in Tab. A3, performance gradually improves as more screened seeds are provided. Even with only 50% of the original seed set, our method maintains competitive performance, demonstrating that the diffusion-based structure learning can effectively exploit sparse screened supervision.

E.2 Diffusion Step Ablation

We further investigate the influence of diffusion sampling steps during inference. As shown in Tab. A4, increasing the number of sampling steps gradually improves the segmentation performance, as more iterative denoising steps allow the model to better recover camouflage structures. However, the performance gain becomes marginal after 10 steps, indicating that excessive sampling steps

Table A4: Ablation study on the number of diffusion sampling steps during inference.

Sampling Steps	MAE↓	S_m ↑	E_m ↑	F_β^w ↑
1	0.034	0.829	0.900	0.732
3	0.032	0.835	0.906	0.739
5	0.031	0.840	0.910	0.742
10	0.030	0.843	0.914	0.744
20	0.030	0.844	0.915	0.744

Table A5: Ablation study of different boundary perturbation strategies on COD10K.

Perturbation Type	**MAE↓**	$\mathbf{S}_m$↑	$\mathbf{E}_m$↑	$\mathbf{F}_\beta^w$↑
None	0.037	0.805	0.877	0.716
Erosion	0.036	0.807	0.878	0.716
Dilation	0.036	0.808	0.879	0.717
Random Morphology	0.036	0.808	0.880	0.718
Contour + Morphology	0.036	0.809	0.882	0.719

Table A6: Comparison between the diffusion-based segmentation head and a conventional decoder with parameter statistics.

Segmentation Head	**Params**	**MAE↓**	$\mathbf{S}_m$↑	$\mathbf{E}_m$↑	$\mathbf{F}_\beta^w$↑
Conventional Decoder	4.99M	0.039	0.796	0.866	0.708
Diffusion Head	7.80M	0.036	0.809	0.882	0.719

provide limited additional benefits. Therefore, we adopt 10 sampling steps as a balance between accuracy and computational efficiency.

E.3 Boundary Perturbation Ablation

We further evaluate the effectiveness of the proposed boundary perturbation strategy $\mathcal{J}(\cdot)$. As shown in Tab. A5, directly training the diffusion head without boundary perturbation leads to inferior performance, demonstrating the importance of introducing structural uncertainty during diffusion learning. Simple morphological perturbations, including erosion and dilation, provide limited improvements by modeling only one type of boundary variation. In contrast, our complete perturbation strategy, which combines contour-level perturbation and local morphological operations, achieves the best performance by generating more diverse boundary-uncertain structural conditions.

F Parameter Comparison with Conventional Decoder

To investigate whether the performance gain of the diffusion-based segmentation head mainly comes from increased model capacity, we compare it with a conventional segmentation decoder under a similar training setting. As shown in Tab. A6, the diffusion head achieves consistent improvements over the standard decoder, despite only introducing a moderate increase in parameters. This indicates that the performance gain is not solely attributed to additional parameters, but benefits from the distributional modeling capability of diffusion-based learning for camouflage structures.