PHYSCORR: RELIABILITY-CONTROLLED CORRECTION FOR VISIBLE-INFRARED CAMOUFLAGED OBJECT SEGMENTATION IN UNSEEN ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visible–infrared camouflaged object segmentation (VICOS) is vulnerable in unseen environments, where the two modalities often suffer asymmetric degradation. Existing fusion methods rely on statistical complementarity and may propagate corrupted modality cues, while visual structures can be overwritten by unconstrained physical priors. To address these limitations, we propose PhysCorr, a reliability-controlled physical correction framework that selectively refines rather than replaces visual evidence. Specifically, imaging conditions are factorized into structured degradation states and target-invariant properties. Shared hierarchical features establish a stable visual reference, while residual correction is generated under physical-state modulation. A reliability-conditioned intervention mechanism integrates cross-modal agreement, branch reliability, spatial uncertainty, and boundary evidence to determine when, where, in which direction, and to what extent correction is applied. Counterfactual invariant learning aligns target representations and invariant reasoning states between observed conditions and clean counterfactuals, suppressing environment-specific shortcuts. Consequently, stable predictions are preserved under conventional conditions, while physical correction is adaptively strengthened in adverse environments. Experiments under conventional and cross-environment settings show that PhysCorr improves source-domain segmentation accuracy and robustness to unseen asymmetric degradation.

1 INTRODUCTION

Camouflaged object segmentation aims to identify objects that closely resemble their surroundings, where low foreground–background contrast often causes weak boundaries, low saliency, and ambiguous structures. Existing methods improve segmentation through distraction mining Mei et al. (2021), uncertainty modeling Li et al. (2021), relational reasoning Zhai et al. (2021), boundary reconstruction He et al. (2023), and global context modeling Huang et al. (2023); Hu et al. (2023). Recent diffusion models and foundation-model adaptation further enhance structural recovery and generalization Chen et al. (2024); Liu et al. (2025b). Nevertheless, most approaches rely primarily on visible appearance and assume sufficiently reliable visual cues at inference.

VICOS uses thermal observations to complement visible perception. Visible images provide texture, color, and boundary cues, while infrared images capture thermal responses that can remain informative under poor illumination or low visible contrast. Existing methods exploit this complementarity through hierarchical fusion, cross-modal attention, confidence estimation, and boundary refinement Tang et al. (2025); Sun et al. (2023); Wang et al. (2024b); Liu et al. (2025a); Jiang et al. (2022), but typically assume stable cross-modal reliability across environments. This assumption breaks under unseen physical conditions because visible and infrared sensors respond differently to environmental interference. Scattering, glare, low illumination, and occlusion may corrupt visible structures, while thermal contrast convergence, insulation effects, and pseudo-hotspots may degrade infrared observations. As shown in Fig. 1, such degradation is asymmetric and spatially varying, making modality reliability environment- and region-dependent. Conventional fusion may therefore propagate incorrect boundaries, amplify spurious thermal responses, or treat degradation artifacts as complementary cues.

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.

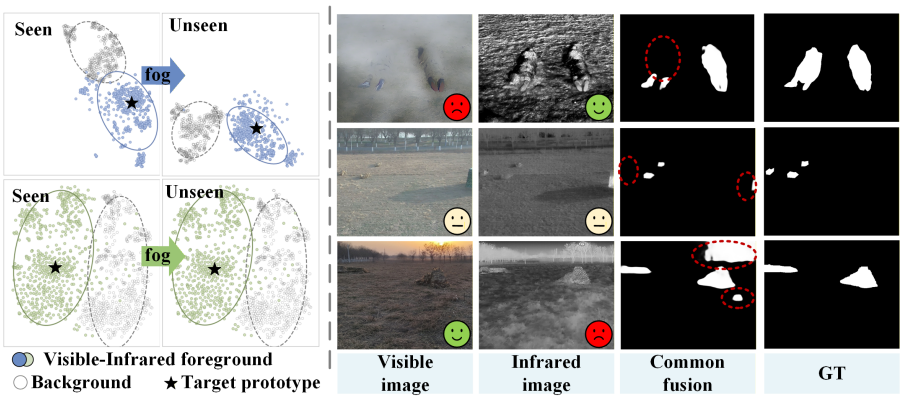

Figure 1: Unseen physical degradations induce modality-specific feature drift and dynamic cross-modal reliability, leading to failures of conventional RGB-T fusion. The left panel analyzes the feature shifts of visible and infrared modalities, while the right panel shows that treating both modalities as equally reliable propagates corrupted observations and causes segmentation failures.

Physical priors can relate environmental conditions to modality-specific degradation and have been explored in adverse-weather restoration and multimodal perception Li et al. (2019; 2020); Zhu et al. (2023); Li et al. (2024b;a); Jiang et al. (2026). However, existing physics-guided methods often treat inferred physical states as reliable guidance. Since these states are estimated from corrupted observations, unconditional intervention may introduce erroneous corrections and overwrite reliable structures. The key challenge is thus to determine when, where, and how strongly physical knowledge should intervene according to the reliability of visual and physical evidence.

To this end, we propose PhysCorr, a reliability-controlled physical correction framework for VICOS in unseen environments. PhysCorr builds a stable visual reference from hierarchical visible and infrared features and models imaging conditions with structured degradation states and target-invariant properties. Instead of producing an independent physics-guided prediction, these states modulate residual correction over the visual reference. A reliability-conditioned intervention mechanism integrates cross-modal agreement, branch reliability, spatial uncertainty, and boundary evidence to control the correction's support and magnitude. Counterfactual invariant learning further aligns target representations and invariant reasoning states between observed and degradation-removed conditions, reducing environment-specific shortcuts.

The main contributions are summarized as follows:

- We propose PhysCorr, which formulates visible–infrared fusion as reliability-controlled physical correction over a stable visual reference for robust segmentation under unseen asymmetric degradation.
- We introduce structured physical state modeling with reliability-conditioned intervention to selectively control the location and strength of physical correction while preserving reliable visual structures.
- We develop counterfactual invariant learning to enforce consistent target representations across observed and counterfactual conditions, improving both in-domain accuracy and cross-environment generalization.

2 Related Work

2.1 Camouflaged Object Segmentation

Early COS methods established benchmarks and focused on foreground–background discrimination Le et al. (2019); Fan et al. (2020). Later studies improved structural recovery through distraction mining Mei et al. (2021), uncertainty reasoning Li et al. (2021), relational modeling Zhai et al. (2021), boundary reconstruction He et al. (2023), and multi-scale context learning Huang et al.

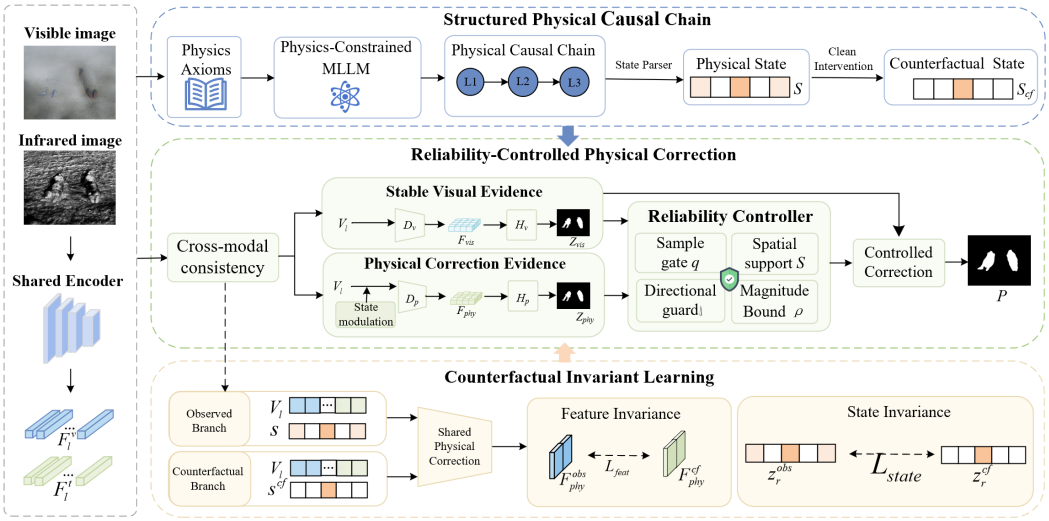

Figure 2: Overall framework of PhysCorr.

(2023); Hu et al. (2023). Recent methods further use diffusion models, weak supervision, prompt learning, and foundation-model adaptation to improve mask generation and reduce annotation dependence Chen et al. (2024); Liu et al. (2025b).

Visible–infrared methods exploit thermal cues through hierarchical fusion, cross-modal interaction, confidence-aware integration, and boundary refinement Tang et al. (2025); Sun et al. (2023); Wang et al. (2024b); Liu et al. (2025a); Jiang et al. (2022). However, they largely rely on statistical complementarity and overlook environment-dependent modality reliability. PhysCorr instead models reliability dynamically to regulate physical correction under asymmetric degradation.

2.2 Physics-Guided Learning

Physics-guided learning introduces image formation, atmospheric degradation, illumination, geometry, and sensor characteristics into visual models, reducing reliance on purely statistical correlations. Physical priors have been widely explored in adverse-weather restoration through scattering, rain formation, transmission, and degradation-specific modeling Li et al. (2019); Halder et al. (2019); Li et al. (2020); Shi et al. (2021); Valanarasu et al. (2022); Zhu et al. (2023). Recent studies further combine physical modeling with multimodal fusion and perception Li et al. (2024b;a), including modality-specific optical and thermal processes in visible–infrared fusion Jiang et al. (2026).

However, most methods use physical knowledge to enhance images or fused features, with limited control over its effect on high-level predictions. When inferred physical states are unreliable, unconditional guidance may conflict with visual evidence. Vision–language models can generate semantic descriptions and reasoning chains Zhang et al. (2024), and recent frameworks extend them to physical scene understanding Balazadeh et al. (2025), but such free-form reasoning is global and spatially coarse for dense segmentation. PhysCorr instead converts physical reasoning into structured states and uses reliability to control residual correction over a stable visual reference, while counterfactual learning suppresses degradation-dependent shortcuts without replacing reliable visual evidence.

3 Method

3.1 Overall Framework

Direct visible–infrared fusion becomes unreliable when the two modalities undergo asymmetric degradation, while inaccurate physical priors may corrupt visual structures that have already been

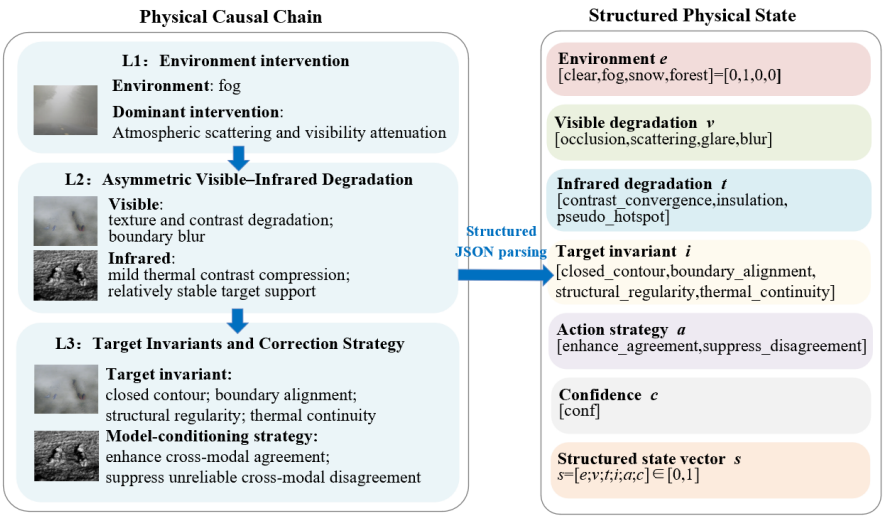

Figure 3: Physical causal chain generation and structured state parsing for fog.

correctly identified. To address these issues, PhysCorr treats physical knowledge as a reliability-controlled residual correction rather than an independent prediction source.

Given a registered visible–infrared pair (I^v, I^t), a shared encoder extracts multi-scale features F_l^v, F_l^t. As shown in Fig. 2, PhysCorr has three components: structured physical prior generation models degradation states and target invariants; reliability-controlled physical correction builds a stable visual reference and applies physics-conditioned residual correction; and counterfactual invariant learning preserves target consistency across environments.

3.2 Structured Physical Causal Chain Generation

Free-form VLM reasoning is unsuitable for direct dense prediction. We use frozen Qwen2.5-VL-7B-Instruct with a fixed prompt to generate a structured physical specification from (I^v, I^t):

$$C = G_{\text{Qwen}}(I^v, I^t) = \{e, d^v, d^t, i, a, c\}, \tag{1}$$

where e is the environment, d^v and d^t are visible and thermal degradations, i denotes target invariants, a the conditioning strategy, and c the confidence.

A deterministic parser maps C to an 18-D physical state:

$$s = \Pi(C) = [s_e; s_v; s_t; s_i; s_a; c] \in [0, 1]^{18}, \tag{2}$$

with dimensions $(4, 4, 3, 4, 2, 1)$. The confidence c is also fed to the reliability controller, and s is encoded as $z_s = \phi_s(s)$.

Since degradation changes observations rather than target identity, we construct a degradation-neutralized counterfactual state during training:

$$s^{\text{cf}} = [s_e^{\text{clear}}; \mathbf{0}_v; \mathbf{0}_t; s_i; s_a; c], \tag{3}$$

where $s_e^{\text{clear}} = [1, 0, 0, 0]$. This intervenes only on degradation-related variables without synthesizing a clean image.

3.3 Reliability-Controlled Physical Correction

Direct fusion may mix complementary cues with degradation artifacts. We therefore build a state-free visual reference and use physical priors only for controlled correction.

3.3.1 Stable Visual Evidence.

Since the inferred physical state may be unreliable, the evidence branch remains independent of physical guidance. At scale l, normalized edge responses E_l^v and E_l^t are used to estimate geometric

agreement: $G_l = 1 - |E_l^v - E_l^t|, \quad M_l = \sigma(2G_l - 1)$, where M_l highlights regions supported by both modalities. Agreement-guided aggregation gives

$$V_l = \mathcal{A}_{\text{vis}}(F_l^v, F_l^t; M_l). \tag{4}$$

The stable reference is decoded as

$$F_{\text{vis}} = D_{\text{vis}}(\{V_l\}_{l=1}^L), \quad Z_{\text{vis}} = H_{\text{vis}}(F_{\text{vis}}), \quad P_{\text{vis}} = \sigma(Z_{\text{vis}}) \tag{5}$$

where D_{vis} and H_{vis} are the state-free decoder and prediction head, and Z_{vis} serves as the stable prediction anchor.

3.3.2 Physics-Conditioned Correction Evidence.

The structured physical state z_s modulates each-scale feature via residual FiLM:

$$\widetilde{V}_l = (1 + \gamma_l(z_s)) \odot V_l + \beta_l(z_s). \tag{6}$$

The modulated features are processed by the physics-conditioned decoder and prediction head to obtain F_{phy}, Z_{phy}, and P_{phy}. This residual design preserves visual evidence while enabling state-dependent enhancement or suppression.

3.3.3 Reliability-Conditioned Intervention.

The discrepancy between the two branches may contain both useful corrections and state-induced errors. We first define a bounded signed residual: $\Delta = \tanh(Z_{\text{phy}} - \text{sg}(Z_{\text{vis}}))$, where $\text{sg}(\cdot)$ denotes stop-gradient. Positive and negative residuals respectively recover missed responses and suppress false ones. Sample-level reliability is estimated from branch features, visual uncertainty, branch disagreement, the physical state, and its confidence:

$$q = \sigma(G_q([\text{GAP}(F_{\text{vis}}); \text{GAP}(F_{\text{phy}}); u; d; z_s; c])), \tag{7}$$

where $u = \text{GAP}(1 - |2P_{\text{vis}} - 1|)$ and $d = \text{GAP}(|P_{\text{phy}} - P_{\text{vis}}|)$. G_q is a two-layer MLP producing scalar reliability q, with c provided as an explicit confidence bypass. A spatial support map localizes uncertain, boundary-sensitive, and cross-modally inconsistent regions:

$$B = \text{Norm}(|\nabla_x P_{\text{vis}}| + |\nabla_y P_{\text{vis}}|), \qquad A = 1 - \frac{1}{L}\sum_{l=1}^{L} \text{Up}(M_l), \tag{8}$$

where B is boundary evidence and A denotes multi-scale cross-modal inconsistency. The support map is

$$S = \text{clip}\left(\rho_0 + \lambda_u U + \lambda_b B + \lambda_d D + \lambda_a A, 0, 1\right), \tag{9}$$

where $U = 1 - |2P_{\text{vis}} - 1|$ and $D = |P_{\text{phy}} - P_{\text{vis}}|$ denote pixel-wise uncertainty and branch disagreement. The final prediction is

$$Z_{\text{out}} = Z_{\text{vis}} + \rho q S \odot \Delta, \qquad P = \sigma(Z_{\text{out}}), \tag{10}$$

where ρ bounds the correction magnitude. Since $q, S \in [0, 1]$ and $|\Delta| \leq 1$, $|Z_{\text{out}} - Z_{\text{vis}}| \leq \rho$. Thus, q, S, and Δ control when, where, in which direction, and how strongly physical correction is applied.

3.4 Counterfactual Invariant Learning

The physics-conditioned branch may exploit weather or modality-specific artifacts as shortcuts. We enforce invariance by removing degradation variables while preserving target-related information.

During training, the observed state $s^{\text{obs}} = s$ and counterfactual state s^{cf} are processed by the shared physics-conditioned branch, yielding $(F_{\text{phy}}^{\text{obs}}, z_r^{\text{obs}})$ and $(F_{\text{phy}}^{\text{cf}}, z_r^{\text{cf}})$. We impose target-aware feature and reasoning-state consistency:

$$\mathcal{L}_{\text{feat}} = \frac{1}{|\Omega_Y|} \sum_{p \in \Omega_Y} \left\| F_{\text{phy}}^{\text{obs}}(p) - \text{sg}\left(F_{\text{phy}}^{\text{cf}}(p)\right) \right\|_1 \tag{11}$$

$$\mathcal{L}_{\text{state}} = \left\| z_r^{\text{obs}} - \text{sg}(z_r^{\text{cf}}) \right\|_2^2 \tag{12}$$

where $\Omega_Y = \{p \mid Y(p) = 1\}$ is the target region. Only the invariant reasoning embedding z_r is aligned, while the degradation embedding z_d remains free to vary.

3.5 LEARNING OBJECTIVE

To preserve the visual pathway as the prediction anchor, we supervise both the stable visual prediction and the final corrected output, while the physics-conditioned pathway is optimized only through bounded residual correction and invariant learning.

For a prediction P', the segmentation loss is

$$\mathcal{L}_{\text{seg}}(P', Y) = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{bd}}. \tag{13}$$

Since correction is applied in logit space, we supervise the actual bounded update $R = \rho q S \odot \Delta$. Let $Z_Y = \text{logit}((1-2\epsilon)Y + \epsilon)$ and $R^* = \text{clip}(Z_Y - \text{sg}(Z_{\text{vis}}), -\rho, \rho)$. We define

$$\mathcal{L}_{\text{res}} = \frac{1}{HW} \sum_p w_p \, \text{SmoothL1}(R(p), R^*(p)). \tag{14}$$

where w_p assigns larger weights to uncertain and boundary pixels. Importantly, Z_{phy} is never directly supervised toward the ground-truth mask; it only contributes through the bounded residual R relative to the detached visual anchor.

The counterfactual objective is $\mathcal{L}_{\text{cf}} = \mathcal{L}_{\text{feat}} + \mathcal{L}_{\text{state}}$. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{seg}}(P_{\text{vis}}, Y) + \mathcal{L}_{\text{seg}}(P, Y) + \lambda_{\text{res}}\mathcal{L}_{\text{res}} + \lambda_{\text{cf}}\mathcal{L}_{\text{cf}}. \tag{15}$$

We first warm up the state-free visual pathway and then activate physical correction and invariant learning, so that residual updates are learned upon a stable visual reference.

4 EXPERIMENT

4.1 EXPERIMENTAL SETTINGS

Datasets. To evaluate both conventional segmentation performance and cross-environment generalization, experiments are conducted on two VICOS datasets. VIACLiu et al. (2025a) is a manually annotated visible–infrared camouflaged object dataset containing 1500 registered image pairs, 1200 pairs are used for training and the remaining 300 pairs are used for testing under conventional environments. EVIACLiu et al. (2026) extends VIAC with three additional environments, including fog, forest, and snow, and contains 6000 image pairs in total. We use the 4500 extended-environment pairs, with 1,500 pairs per environment, exclusively for cross-environment testing.

Evaluation Metrics. In this work, performance is evaluated using four metrics: S-measure(S_m) Fan et al. (2017) for evaluating structural information, weighted F-measure(F_β^w) Margolin et al. (2014) for measuring precision and recall in the predicted maps, mean absolute error (*MAE*) Perazzi et al. (2012) for calculating element-wise differences between predictions and ground truth, and adaptive E-measure(E_m) Fan et al. (2018) for evaluating both local and global information.

Implementation details. Our framework is built upon PVTv2-B4 and implemented in PyTorch. The model is trained on an NVIDIA RTX PRO 6000 GPU for 100 epochs with a batch size of 10. We employ AdamW with an initial learning rate of 1×10^{-4} and cosine decay. All visible–infrared pairs are resized to 384×384. Qwen2.5-VL-7B-Instruct is used only to generate structured physical causal chains and remains frozen throughout training.

4.2 RESULTS AND DATA ANALYSIS

Performance under Conventional Environments. Table 1 reports quantitative comparisons on the VIAC dataset under conventional environments. PhysCorr consistently achieves the best overall performance, improving F_β^w by 0.44% over the strongest RGB–T baseline. Although physical correction is designed to address unseen physical degradations, these results indicate that the proposed reliability-controlled mechanism does not interfere with reliable observations under normal conditions. Instead, it preserves stable visual representations while introducing complementary physical guidance only when beneficial, thereby maintaining strong source-domain performance before evaluating cross-environment generalization.

Performance under Unseen Environments Table 2 reports quantitative comparisons on the EVIAC benchmark under three representative unseen environments, including fog, forest, and

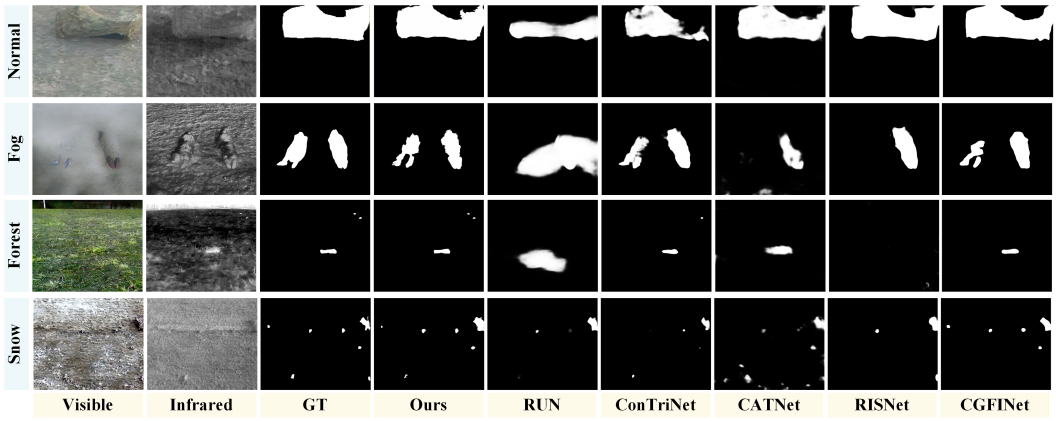

Figure 4: Qualitative comparison on the VIAC and EVIAC datasets.

Methods	Tasks	VIAC-test			
		$S_m \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$	$E_m \uparrow$
FEDER He et al. (2023)	RGB COD	0.9099	0.8792	0.0155	0.9502
ZoomNeXtPang et al. (2024)	RGB COD	0.9030	0.8630	0.0160	0.9590
PlantCamo Yang et al. (2025)	RGB COD	0.9262	0.9186	0.0107	0.9827
RUNHe et al. (2025)	RGB COD	0.9230	0.9042	0.0128	0.9747
LSNet Zhou et al. (2023)	RGB-T SOD	0.9154	0.8947	0.0138	0.9660
SPNet Zhang et al. (2023b)	RGB-T SOD	0.9288	0.9233	0.0105	0.9798
SACNet Wang et al. (2024a)	RGB-T SOD	0.9274	0.9156	0.0111	0.9818
ConTriNet Tang et al. (2025)	RGB-T SOD	0.9024	0.8878	0.0182	0.9595
HIDANet Wu et al. (2023a)	RGB-D SOD	0.9268	0.9130	0.0122	0.9705
CATNet Sun et al. (2023)	RGB-D SOD	0.8948	0.8409	0.0190	0.9262
DCMF Wang et al. (2022)	RGB-D SOD	0.9247	0.8966	0.0132	0.9481
MoADNet Jin et al. (2022)	RGB-D SOD	0.8663	0.8255	0.0262	0.9447
XMSNet Wu et al. (2023c)	RGB-D COD	0.9268	0.9130	0.0122	0.9705
DaCOD Wang et al. (2023)	RGB-D COD	0.9183	0.8982	0.0139	0.9708
PopNet Wu et al. (2023b)	RGB-D COD	0.9255	0.9109	0.0117	0.9714
RISNet Wang et al. (2024b)	RGB-D COD	0.9292	0.9214	0.0105	0.9791
CGFINet Liu et al. (2025a)	RGB-T COD	0.9322	0.9254	0.0105	0.9862
HCRNet Liu et al. (2026)	RGB-T COD	0.9323	0.9233	0.0103	0.9825
Ours	RGB-T COD	**0.9350**	**0.9298**	**0.0100**	**0.9863**

Table 1: Quantitative comparison with state-of-the-art visible and multimodal COS methods on the VIAC dataset under conventional environments.

snow. We compare PhysCorr with representative RGB-based COD, RGB–D COD, and RGB–T COD/SOD methods. Despite the substantially different degradation characteristics across environments, PhysCorr consistently achieves the best performance in all three settings, demonstrating strong cross-environment generalization. Among the evaluated environments, fog is the most challenging because atmospheric scattering simultaneously weakens visible structural cues and reduces infrared thermal contrast, resulting in severe asymmetric modality degradation. Forest scenes present intermediate difficulty due to increased background complexity and stronger camouflage, while snow yields the best performance because reliable thermal cues remain largely preserved and effectively compensate for degraded visible observations. These results validate that explicitly modeling environment-dependent modality reliability enables PhysCorr to maintain robust segmentation performance across diverse unseen physical environments.

Qualitative Comparison Figure 4 presents qualitative comparisons under both conventional and unseen environments. Under conventional conditions, although most methods produce reasonable predictions, they still suffer from incomplete object structures or blurred boundaries in challenging regions. PhysCorr consistently preserves more complete targets and finer structural details without introducing unnecessary corrections. The superiority of PhysCorr becomes more evident under unseen environments. Fog causes severe degradation of both visible structures and thermal contrast, forest increases foreground–background ambiguity due to complex textures, and snow mainly

Methods	EVIAC-fog				EVIAC-forest				EVIAC-snow			
	$S_m \uparrow$	$F_\beta^w \uparrow$	MAE$\downarrow$	$E_m \uparrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	MAE$\downarrow$	$E_m \uparrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	MAE$\downarrow$	$E_m \uparrow$
RUN He et al. (2025)	0.6866	0.5556	0.1298	0.7496	0.8587	0.8059	0.0303	0.9245	0.9458	0.9395	0.0084	0.9834
ConTriNet Tang et al. (2025)	0.8335	0.7808	0.0448	0.9156	0.9204	0.9068	0.0128	0.9748	0.9187	0.9110	0.0142	0.9708
CATNet Sun et al. (2023)	0.8124	0.7060	0.0394	0.8351	0.8535	0.7646	0.0289	0.8879	0.8968	0.8397	0.0172	0.9205
RISNet Wang et al. (2024b)	0.8227	0.7573	0.0440	0.8938	0.9088	0.8880	0.0133	0.9537	0.9112	0.8927	0.0127	0.9593
CGFINet Liu et al. (2025a)	0.8726	0.8625	0.0298	0.9458	0.9372	0.9344	0.0091	0.9866	0.9375	0.9367	0.0094	0.9863
Ours	**0.8973**	**0.8907**	**0.0237**	**0.9598**	**0.9494**	**0.9505**	**0.0073**	**0.9894**	**0.9530**	**0.9542**	**0.0068**	**0.9896**

Table 2: Quantitative comparison on the EVIAC dataset under unseen physical environments.

Group	Item	VIAC-test		EVIAC	
		$F_\beta^w \uparrow$	MAE $\downarrow$	$F_\beta^w \uparrow$	MAE $\downarrow$
(a)	Visual baseline	0.9192	0.0117	0.9124	0.0141
	+ Structured state	0.9239	0.0110	0.9223	0.0133
	+ Controlled correction	0.9276	0.0105	0.9276	0.0129
(b)	Free-form chain	0.9246	0.0108	0.9228	0.0132
	w/o degradation states	0.9258	0.0107	0.9248	0.0131
	w/o target invariants	0.9269	0.0106	0.9264	0.0130
(c)	Direct logits fusion	0.9244	0.0108	0.9211	0.0134
	w/o sample gate q	0.9259	0.0107	0.9247	0.0132
	w/o spatial support S	0.9268	0.0106	0.9259	0.0131
	w/o directional guard	0.9279	0.0104	0.9285	0.0128
	w/o magnitude bound ρ	0.9283	0.0102	0.9289	0.0128
(d)	No counterfactual learning	0.9266	0.0106	0.9262	0.0131
	w/o $\mathcal{L}_{\text{feat}}$	0.9276	0.0104	0.9276	0.0129
	w/o $\mathcal{L}_{\text{state}}$	0.9285	0.0102	0.9292	0.0128
	PhysCorr	**0.9298**	**0.0100**	**0.9318**	**0.0126**

Table 3: Ablation studies of PhysCorr on VIAC-test and EVIAC.

affects visual appearance while preserving relatively reliable thermal cues. Existing methods therefore exhibit missed detections, false positives, or boundary distortions under different environments. In contrast, PhysCorr consistently produces more accurate and structurally complete segmentation results by selectively suppressing unreliable modality responses and introducing physical guidance only when it provides reliable corrective information. These observations further validate the effectiveness of reliability-controlled physical correction for robust cross-environment segmentation.

4.3 Ablation studies

Overall component analysis. Table 3(a) presents the overall contribution of the three proposed components. Starting from the visual baseline, each module consistently improves performance on both VIAC and EVIAC, while the gain on EVIAC is larger. This observation indicates that the proposed design primarily enhances robustness under unseen physical degradations while preserving strong in-domain performance. Structured physical states improve environment awareness, reliability-controlled correction suppresses unreliable modality responses, and counterfactual learning further reduces environment-specific bias.

Structured physical causal chain generation. Table 3(b) validates the effectiveness of structured physical prior generation. Replacing the structured state with free-form physical reasoning leads to the largest performance degradation, especially on EVIAC, indicating that explicit state factorization provides more stable physical representations than unconstrained language descriptions. Removing degradation states weakens the ability to distinguish modality-specific corruption, while removing target invariants causes environmental responses to become entangled with target identity. Both components contribute more noticeably under unseen environments, confirming the importance of explicitly disentangling degradation and invariant properties.

Reliability-controlled physical correction. Table 3(c) evaluates the reliability-controlled correction strategy. Directly fusing visual and physical predictions results in the largest performance drop, demonstrating that indiscriminate physical correction may overwrite reliable visual evidence. Re-

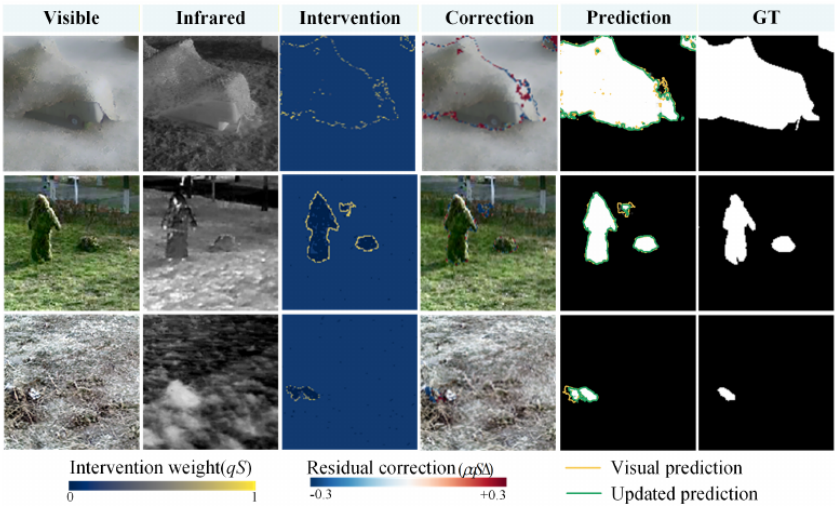

Figure 5: Visualization of reliability-controlled physical correction under representative unseen environments. The intervention weight identifies regions requiring physical intervention, while the residual correction suppresses unreliable responses or enhances missing structures. The updated predictions show that PhysCorr selectively corrects uncertain regions while preserving stable visual evidence, producing results closer to the GT under fog, forest, and snow conditions.

moving the sample gate or spatial support also noticeably degrades performance, indicating that determining when and where physical correction should be applied is crucial. In comparison, removing directional or magnitude constraints causes relatively smaller degradation, suggesting that reliable activation is more important than the specific correction form.

Figure 5 further provides a visual explanation of this behavior. The intervention weights concentrate primarily on degraded boundaries, ambiguous target regions, and locally inconsistent cross-modal responses, while confident regions remain largely inactive. Guided by these weights, the residual correction selectively suppresses false responses and enhances missing structures instead of modifying the entire prediction. Across fog, forest, and snow conditions, the corrected contours move closer to the ground truth while most reliable structures remain unchanged. These observations are consistent with the ablation results and confirm that the main benefit of PhysCorr arises from reliability-aware activation of physical correction rather than unconditional physical guidance.

Counterfactual invariant learning. Table 3(d) studies the contribution of counterfactual invariant learning. Removing the entire counterfactual branch significantly reduces cross-environment performance, confirming its role in suppressing environment-specific shortcuts. Among individual objectives, feature alignment contributes the most by preserving consistent target representations across observed and counterfactual conditions, while state alignment further stabilizes physical reasoning. Their combination produces the strongest robustness under unseen environments, demonstrating that invariant learning effectively complements reliability-controlled physical correction.

5 Conclusion

We presented PhysCorr, a reliability-controlled physical correction framework for VICOS in unseen environments. Physical degradations mainly alter observations rather than target identity, leading to environment-dependent reliability across modalities. Accordingly, PhysCorr models degradation as structured states, selectively applies physical correction based on cross-modal reliability, and promotes environment-invariant representations through counterfactual learning. Extensive experiments on conventional and cross-environment benchmarks show that PhysCorr consistently improves segmentation accuracy and generalization under unseen physical conditions. More broadly, our results suggest that physical knowledge should act as adaptive corrective evidence rather than replace reliable visual observations, offering a principled paradigm for robust multimodal perception in complex environments.

AI USE STATEMENT

In this work, we used generative AI tools for research ideation and execution, including providing feedback on research methodology, experimental design, and implementation details. We have not used generative AI tools for generating synthetic datasets or proving mathematical claims, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for retrieval and discovery of relevant literature and for editing and polishing author-written text to improve clarity and readability. We have reviewed all AI-assisted work. AI-assisted research ideas and implementation suggestions were independently examined and validated by the authors through implementation and experiments, and AI-suggested references were manually checked against the original sources. All AI-assisted writing was reviewed and revised by the authors to ensure technical accuracy and consistency with the actual methodology and experimental results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Vahid Balazadeh, Mohammadmehdi Ataei, Hyunmin Cheong, Amir Hosein Khasahmadi, and Rahul G. Krishnan. Physics context builders: A modular framework for physical reasoning in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7318–7328, 2025.

Geng Chen, Si-Jie Liu, Yu-Jia Sun, Ge-Peng Ji, Ya-Feng Wu, and Tao Zhou. Camouflaged object detection via context-aware cross-level fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6981–6993, 2022.

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, 2024. doi: 10.1609/aaai.v38i2.27890.

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. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 698–704, 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, 2020.

Dengping Fan, Gepeng Ji, Mingming Cheng, and Ling Shao. Concealed object detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6024–6042, 2021.

Shirsendu Sukanta Halder, Jean-Francois Lalonde, and Raoul de Charette. Physics-based rendering for improving robustness to rain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10203–10212, 2019.

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, 2023.

Chunming He, Rihan Zhang, Fengyang Xiao, Chenyu Fang, Longxiang Tang, Yulun Zhang, Linghe Kong, Deng-Ping Fan, Kai Li, and Sina Farsiu. Run: Reversible unfolding network for concealed object segmentation. *ICML*, 2025.

Xiaobin Hu, Shuo Wang, Xuebin Qin, Hang Dai, Wenqi Ren, Donghao Luo, Ying Tai, and Ling Shao. High-resolution iterative feedback network for camouflaged object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 881–889, 2023. doi: 10.1609/aaai.v37i1.25167.

Xihang Hu, Xiaoli Zhang, Fasheng Wang, Jing Sun, and Fuming Sun. Efficient camouflaged object detection network based on global localization perception and local guidance refinement. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):5452–5465, 2024.

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*, pp. 5557–5566, 2023.

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*, pp. 4713–4722, 2022.

Haiyang Jiang, Huiqin Zhang, Yanduo Zhang, Jiayi Ma, Junjun Jiang, and Huabing Zhou. PhyFusion: Physics-aware infrared and visible image fusion via modality-specific physical priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 4993–5002, 2026.

Xiurong Jiang, Lin Zhu, Yifan Hou, and Hui Tian. Mirror complementary transformer network for RGB-thermal salient object detection. *arXiv preprint arXiv:2207.03558*, 2022.

Xiao Jin, Kang Yi, and Jing Xu. Moadnet: Mobile asymmetric dual-stream networks for real-time and lightweight rgb-d salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7632–7645, 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. doi: 10.1016/j.cviu.2019.04.006.

Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10071–10081, 2021.

Haoyuan Li, Qi Hu, You Yao, Yang Kailun, and Peng Chen. Cfmw: Cross-modality fusion mamba for multispectral object detection under adverse weather conditions. *arXiv preprint arXiv:2404.16302*, 2024a.

Ruoteng Li, Loong-Fah Cheong, and Robby T. Tan. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1633–1642, 2019.

Ruoteng Li, Robby T. Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Xilai Li, Wuyang Liu, Xiaosong Li, Fuqiang Zhou, Huafeng Li, and Feiping Nie. All-weather multi-modality image fusion: Unified framework and 100k benchmark. *arXiv preprint arXiv:2402.02090*, 2024b.

Cheng Liu, Zheng Wang, Xinyu Yan, Meijun Sun, and Qinghua Hu. Visible-infrared camouflaged object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025a.

Cheng Liu, Zheng Wang, Yahong Han, Meijun Sun, and Qinghua Hu. Hierarchical cross-modal reasoning for visible-infrared camouflaged object detection. *IEEE Transactions on Multimedia*, pp. 1–12, 2026.

Jiaming Liu, Linghe Kong, and Guihai Chen. Improving sam for camouflaged object detection via dual stream adapters. In *Proceedings of the IEEE International Conference on Computer Vision*, 2025b.

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*, pp. 8772–8781, 2021.

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*, pp. 2160–2170, 2022.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE transactions on pattern analysis and machine intelligence*, 46(12):9205–9220, 2024.

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.

Zheng Shi, Ethan Tseng, Mario Bijelic, Werner Ritter, and Felix Heide. Zeroscatter: Domain transfer for long distance imaging and vision through scattering media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3476–3486, 2021.

Ze Song, Xudong Kang, Xiaohui Wei, Haibo Liu, Renwei Dian, and Shutao Li. Fsnet: Focus scanning network for camouflaged object detection. *IEEE Transactions on Image Processing*, 32: 2267–2278, 2023.

Fuming Sun, Peng Ren, Bowen Yin, Fasheng Wang, and Haojie Li. Catnet: A cascaded and aggregated transformer network for rgb-d salient object detection. *IEEE Transactions on Multimedia*, 26:2249–2262, 2023.

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

Yanguang Sun, Chunyan Xu, Jian Yang, Hanyu Xuan, and Lei Luo. Frequency-spatial entanglement learning for camouflaged object detection. In *European Conference on Computer Vision*, pp. 343–360. Springer, 2024a.

Yanguang Sun, Hanyu Xuan, Jian Yang, and Lei Luo. Glconet: Learning multisource perception representation for camouflaged object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2024b.

Yujia Sun, Shuo Wang, Chenglizhao Chen, and Tian-Zhu Xiang. Boundary-guided camouflaged object detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 1335–1341, 2022.

Hao Tang, Zechao Li, Dong Zhang, Shengfeng He, and Jinhui Tang. Divide-and-conquer: Confluent triple-flow network for rgb-t salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1958–1974, 2025.

Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2353–2363, 2022.

Fengyun Wang, Jinshan Pan, Shoukun Xu, and Jinhui Tang. Learning discriminative cross-modality features for rgb-d saliency detection. *IEEE Transactions on Image Processing*, 31:1285–1297, 2022.

Huizhi Wang, Hui Guo, Xiongli Chai, Baoyang Mu, and Feng Shao. Cognition-inspired dynamic feature integration network for rgb-d and rgb-t salient object detection. *IEEE Transactions on Instrumentation and Measurement*, 2025.

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

Liqiong Wang, Jinyu Yang, Yanfu Zhang, Fangyi Wang, and Feng Zheng. Depth-aware concealed crop detection in dense agricultural scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17201–17211, 2024b.

Qingwei Wang, Jinyu Yang, Xiaosheng Yu, Fangyi Wang, Peng Chen, and Feng Zheng. Depth-aided camouflaged object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3297–3306, 2023.

Zongwei Wu, Guillaume Allibert, Fabrice Meriaudeau, Chao Ma, and Cédric Demonceaux. Hidanet: Rgb-d salient object detection via hierarchical depth awareness. *IEEE Transactions on Image Processing*, 32:2160–2173, 2023a. doi: 10.1109/TIP.2023.3263111.

Zongwei Wu, Danda Pani Paudel, Deng-Ping Fan, Jingjing Wang, Shuo Wang, Cédric Demonceaux, Radu Timofte, and Luc Van Gool. Source-free depth for object pop-out. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1032–1042, 2023b.

Zongwei Wu, Jingjing Wang, Zhuyun Zhou, Zhaochong An, Qiuping Jiang, Cédric Demonceaux, Guolei Sun, and Radu Timofte. Object segmentation by mining cross-modal semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3455–3464, 2023c.

Chenxi Xie, Changqun Xia, Tianshu Yu, and Jia Li. Frequency representation integration for camouflaged object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1789–1797, 2023.

Jinyu Yang, Qingwei Wang, Feng Zheng, Peng Chen, Ale Leonardis, and Deng Ping Fan. Plantcamo: Plant camouflage detection. *CAAI Artificial Intelligence Research*, 4:9150045, 2025.

Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12997–13007, 2021.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024.

Xinyue Zhang, Jiahuan Zhou, Luxin Yan, Sheng Zhong, and Xu Zou. Hunt camouflaged objects via revealing mutation regions. *IEEE Transactions on Information Forensics and Security*, 20: 1836–1851, 2025.

Yi Zhang, Jing Zhang, Wassim Hamidouche, and Olivier Deforges. Predictive uncertainty estimation for camouflaged object detection. *IEEE Transactions on Image Processing*, 32:3580–3591, 2023a.

Zihao Zhang, Jie Wang, and Yahong Han. Saliency prototype for rgb-d and rgb-t salient object detection. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 3696–3705, 2023b.

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3:9150038, 2024.

Wujie Zhou, Yun Zhu, Jingsheng Lei, Rongwang Yang, and Lu Yu. Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images. *IEEE Transactions on Image Processing*, 32:1329–1340, 2023.

Yurui Zhu, Tianyu Wang, Xueyang Fu, Xuanyu Yang, Xin Guo, Jifeng Dai, Yu Qiao, and Xiaowei Hu. Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21747–21758, 2023.

A APPENDIX

A.1 IMPLEMENTATION DETAILS

This section provides additional implementation details that are omitted from the main paper due to space limitations. We mainly describe the dataset protocol, physical prior preparation, training strategy, and overall optimization procedure of PhysCorr.

A.1.1 PHYSICAL PRIOR PREPARATION

The structured physical states are generated once and cached before segmentation training/evaluation. We use the frozen Qwen2.5-VL-7B-Instruct model for state acquisition only; it is not included in the segmentation forward pass. For paired visible–infrared inference, we use near-deterministic sampling with `do_sample=True`, `temperature` $= 10^{-6}$, `repetition_penalty` $= 1.05$, and `max_new_tokens` $= 1024$. For the text-based environment templates used during training preparation, the maximum generation length is 512 tokens.

The model is required to return a predefined JSON specification together with a short `raw_chain_text`. The JSON contains the environment type, optical and thermal degradation variables, cross-modal invariants, decoupling strategies, and confidence. The exact prompts and output schema are provided below. The textual rationale is stored only for interpretation and auditing and is not used as an effective text-conditioning signal by the reported segmentation model.

A deterministic non-learnable parser first extracts the JSON object from the generated response and normalizes all predefined fields. Finite numeric values are clipped to $[0, 1]$. Missing or invalid fields are replaced by the corresponding environment-specific default values. If no recoverable structured dictionary can be obtained, the default physical specification is used and the generation is marked invalid for diagnostic purposes. No learned repair module or additional semantic-consistency model is used.

For each image pair, the cache stores the normalized physical specification, the physical-state vector, its degradation-neutralized counterpart, the confidence value, and the human-readable causal-chain text. During segmentation training and evaluation, these cached quantities are loaded directly. Consequently, Qwen2.5-VL is not invoked inside the segmentation forward pass. For an uncached sample, state generation and parsing are performed once before the resulting state is written back to the cache.

A.1.2 EXACT PROMPT AND OUTPUT SCHEMA

We use a fixed instruction for paired visible-infrared physical-state acquisition. The prompt requires the frozen Qwen2.5-VL-7B-Instruct model to analyze the registered image pair and return a structured physical specification with a short human-readable rationale. No segmentation mask or target annotation is provided to the model.

The instruction follows the template below:

```
Given a registered visible image and infrared image of
the same scene, analyze the physical factors affecting
visible-infrared perception.
Identify the environment condition and estimate
modality-specific degradation, cross-modal target
invariants, model-conditioning strategies, and the
confidence of the inferred physical state.
The environment type must be selected from:  clear, fog,
snow, forest.
For the visible modality, estimate:  occlusion, scattering,
glare, and blur.
For the infrared modality, estimate:  contrast convergence,
insulation, and pseudo hotspot.
For cross-modal target invariants, estimate:  closed
contour, boundary alignment, structural regularity, and
thermal continuity.
```

Algorithm 1 Training and inference procedure of PhysCorr

Input: Visible-infrared pair (I^v, I^t); mask Y during training; cached physical states (s, s^{cf})
Parameter: Encoder E; state encoder ϕ_s; visual branch $\mathcal{V}$; physical correction branch $\mathcal{P}$; reliability controller $\mathcal{R}$; warm-up period T_w
Output: Prediction P and training loss $\mathcal{L}$

Load physical state s and degradation-neutralized state s^{cf}
Extract confidence $c \leftarrow s[-1]$
Encode physical state $z_s \leftarrow \phi_s(s)$
$\{F_l^v\}, \{F_l^t\} \leftarrow E(I^v), E(I^t)$
$(F_v, Z_v) \leftarrow \mathcal{V}(\{F_l^v\}, \{F_l^t\})$
$P_v \leftarrow \mathrm{Sigmoid}(Z_v)$
if training and current epoch $< T_w$ **then**
    $\mathcal{L} \leftarrow \mathcal{L}_{seg}(P_v, Y)$
    return $P_v, \mathcal{L}$
end if
$(F_p, Z_p, z_r) \leftarrow \mathcal{P}(\{F_l^v\}, \{F_l^t\}, z_s)$
$(q, S) \leftarrow \mathcal{R}(F_v, F_p, Z_v, Z_p, z_s, c)$
$\Delta \leftarrow \tanh(Z_p - \mathrm{StopGrad}(Z_v))$
$R \leftarrow \rho q S \odot \Delta$
$Z_{\mathrm{out}} \leftarrow Z_v + R$
$P \leftarrow \mathrm{Sigmoid}(Z_{\mathrm{out}})$
if training **then**
    Encode degradation-neutralized state $z_s^{cf} \leftarrow \phi_s(s^{cf})$
    $(F_p^{cf}, Z_p^{cf}, z_r^{cf}) \leftarrow \mathcal{P}(\{F_l^v\}, \{F_l^t\}, z_s^{cf})$
    Compute invariant consistency: $\mathcal{L}_{cf} \leftarrow \mathrm{Align}(F_p, F_p^{cf}, z_r, z_r^{cf}, Y)$
    $Y_\epsilon \leftarrow (1 - 2\epsilon)Y + \epsilon$
    $Z_Y \leftarrow \mathrm{Logit}(Y_\epsilon)$
    $R^* \leftarrow \mathrm{Clip}(Z_Y - \mathrm{StopGrad}(Z_v), -\rho, \rho)$
    $\mathcal{L}_{res} \leftarrow \frac{1}{HW} \sum_p w_p \, \mathrm{SmoothL1}(R(p), R^*(p))$
    Optimize: $\mathcal{L} \leftarrow \mathcal{L}_{seg}(P_v, Y) + \mathcal{L}_{seg}(P, Y) + \lambda_{res}\mathcal{L}_{res} + \lambda_{cf}\mathcal{L}_{cf}$
    return $P, \mathcal{L}$
else
    return P
end if

```
For model-conditioning strategies, estimate:  enhance
agreement, suppress disagreement, prefer IR coarse
localization, and refine RGB boundary.
All numerical attributes must be represented by continuous
values in [0, 1].  Also provide a concise three-level physical
rationale describing (1) the environmental intervention,
(2) asymmetric visible-infrared degradation, and (3) the
target-preserving response.
Return the physical specification using the predefined JSON
schema.
```

The structured output follows the schema:

```
{
 "env_type":
 "clear|fog|snow|forest",

 "optical_degradation": {
  "occlusion": 0.0,
  "scattering": 0.0,
  "glare": 0.0,
  "blur": 0.0
 },
```

```
  "thermal_degradation": {
   "contrast_convergence": 0.0,
   "insulation": 0.0,
   "pseudo_hotspot": 0.0
  },

  "cross_modal_invariants": {
   "closed_contour": 0.0,
   "boundary_alignment": 0.0,
   "structural_regularity": 0.0,
   "thermal_continuity": 0.0
  },

  "decoupling_strategy": {
   "enhance_agreement": 0.0,
   "suppress_disagreement": 0.0,
   "prefer_ir_coarse": 0.0,
   "refine_rgb_boundary": 0.0
  },

  "confidence": 0.0
 }
```

Among the four fields in `decoupling_strategy`, only `enhance_agreement` and `suppress_disagreement` are used as effective conditioning variables in the reported model. `prefer_ir_coarse` and `refine_rgb_boundary` are retained only as diagnostic metadata. The accompanying three-level rationale is likewise retained for interpretation and auditing and is not used as an effective text-conditioning input to the segmentation network.

For paired-image state inference, generation uses `do_sample=True`, `temperature` $= 10^{-6}$, `repetition_penalty` $= 1.05$, and `max_new_tokens` $= 1024$. Because the temperature is extremely small, we refer to this setting as near-deterministic sampling rather than greedy decoding. For text-based environment-template generation used during training preparation, the maximum generation length is 512 tokens.

A.1.3 Parsing and Caching Rules

The generated response is processed by a deterministic, non-learnable parser. We first search for a fenced JSON block in the generated text. If no fenced block is found, the parser attempts to recover the outermost JSON object delimited by the first valid opening and closing braces. The recovered dictionary is then matched against the predefined schema.

For all numerical attributes, finite values are converted to floating-point numbers and clipped to $[0, 1]$. Missing, malformed, or non-finite fields are replaced by the corresponding environment-specific default values rather than by a universal constant. The environment label is normalized to one of the four supported categories before state construction. No learned repair module or semantic state-conflict detector is applied, and the confidence score is not heuristically reduced when different generated fields appear semantically inconsistent.

If no recoverable structured dictionary can be obtained for paired-image inference, the parser returns the predefined default physical specification and marks the generation as invalid for diagnostic purposes. This validity flag is used only to record state-generation status and is not introduced as an additional conditioning variable. During training-template preparation, a failed generation may instead fall back to the predefined seeded template generation procedure.

For each image pair, the physical-state cache stores the resolved environment type, human-readable causal-chain text, normalized physical specification, physical-state vector, degradation-neutralized state vector, and state confidence. The raw unprocessed Qwen response and parsing-status metadata are not stored in the cache.

Case	Processing rule
Fenced JSON	Extract enclosed JSON object
No fenced JSON	Recover outermost JSON object
Value outside $[0, 1]$	Clip to $[0, 1]$
Missing/invalid field	Environment-specific default
Unrecoverable response	Default specification + invalid flag
Cache hit	Normalize and directly retrieve state
Cache miss	Infer/construct state and cache once

Table 4: Deterministic parsing and cache handling rules.

At test time, the system first queries the physical-state cache. For a cache hit, the cached state is normalized and directly used by the segmentation model; an explicitly specified environment override, when present in the evaluation protocol, is applied after cache retrieval. For a cache miss, an explicit environment override can be used to construct the corresponding deterministic protocol state. Otherwise, Qwen2.5-VL is invoked once on the registered visible-infrared pair, after which the generated response is parsed, normalized, and written to the cache. Consequently, the vision-language model is not invoked inside the segmentation forward pass when a valid cached state is available.

A.2 Training Strategy

PhysCorr adopts a progressive training strategy to stabilize physical correction learning. During the warm-up stage, only the visual branch is optimized to obtain reliable visible–infrared representations. After the warm-up stage, the physical correction branch and reliability controller are activated.

During the full training stage, the model is jointly optimized with segmentation objectives, residual correction constraints, and counterfactual invariant learning. The observed and counterfactual physical states share the same physical correction branch, encouraging the model to preserve target-related information while reducing environment-specific variations.

To improve robustness under unseen environments, we additionally apply physics-guided domain expansion during training. Specifically, physical degradation factors are used to simulate environment-induced variations in visible and infrared modalities, such as contrast attenuation, boundary weakening, thermal response variation, and background interference. The generated samples preserve the original target masks and are only used as training augmentation.

During inference, only the visible–infrared image pair and the cached physical state are required. All training-specific operations, including counterfactual learning and domain expansion, are removed.

The overall procedure is summarized in Algorithm 1.

A.3 Additional Quantitative and Qualitative Comparisons

To further evaluate the effectiveness of PhysCorr, we provide additional quantitative and qualitative comparisons with state-of-the-art camouflage object segmentation methods. Quantitative results are reported on both the conventional VIAC benchmark and the unseen-environment EVIAC benchmark, while qualitative comparisons are presented under representative environmental conditions, including fog, forest, and snow.

A.3.1 Quantitative Comparisons

Table 5 reports the comparison results on the VIAC dataset. PhysCorr achieves the best overall performance among all compared methods. Compared with previous state-of-the-art methods, PhysCorr provides consistent improvements across different evaluation metrics, demonstrating its effectiveness in visible–infrared camouflage segmentation under conventional conditions. The improvement mainly benefits from maintaining stable visual evidence and introducing physical correction only when complementary information is required, which avoids excessive modification of reliable regions.

Methods	VIAC-test			
	$S_m \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$	$E_\phi^a \uparrow$
SINet Fan et al. (2020)	0.9005	0.7859	0.0197	0.9487
PFNet Mei et al. (2021)	0.9247	0.9095	0.0126	0.9309
SINetV2 Fan et al. (2021)	0.9269	0.9137	0.0114	0.9779
BGNet Sun et al. (2022)	0.9266	0.9158	0.0114	0.9320
C2FNetV2 Chen et al. (2022)	0.9209	0.9068	0.0120	0.9294
SegMaR Jia et al. (2022)	0.9273	0.9115	0.0115	0.9732
ZoomNet Pang et al. (2022)	0.9178	0.8962	0.0127	0.9685
FRINet Xie et al. (2023)	0.9302	0.9180	0.0113	0.9788
HitNet Hu et al. (2023)	0.9295	0.9201	0.0112	0.9819
PUENet Zhang et al. (2023a)	0.9210	0.9153	0.0112	0.9822
FSNet Song et al. (2023)	0.9291	0.9181	0.0113	0.9793
FSEL Sun et al. (2024a)	0.9318	0.9210	0.0107	0.9803
GLCONetSun et al. (2024b)	0.9233	0.9065	0.0116	0.9679
BiRefNetZheng et al. (2024)	0.9232	0.9218	0.0106	0.9808
PRNetHu et al. (2024)	0.9308	0.9244	0.0108	0.9807
MRNetZhang et al. (2025)	0.9150	0.8857	0.0149	0.9353
CamoDiffusionSun et al. (2025)	0.9321	0.9158	0.0107	0.9796
Ours	**0.9350**	**0.9298**	**0.0100**	**0.9863**

Table 5: Compare with state-of-the-art COD models on VIAC datasets.

Methods	EVIAC-fog				EVIAC-forest				EVIAC-snow			
	$S_m \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$	$E_\phi^a \uparrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$	$E_\phi^a \uparrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$	$E_\phi^a \uparrow$
SINet Fan et al. (2020)	0.6778	0.4913	0.1014	0.7856	0.8120	0.6468	0.0485	0.8299	0.8320	0.6900	0.0396	0.8562
PFNet Mei et al. (2021)	0.6918	0.5554	0.0974	0.7521	0.9087	0.8816	0.0146	0.9281	0.9128	0.8884	0.0133	0.9267
SINetV2 Fan et al. (2021)	0.7413	0.6423	0.0713	0.8468	0.9181	0.9061	0.0130	0.9668	0.9228	0.9075	0.0108	0.9647
BGNet Sun et al. (2022)	0.7407	0.6362	0.0914	0.8079	0.8696	0.8213	0.0293	0.8997	0.9189	0.8909	0.0134	0.9220
C2FNetV2 Chen et al. (2022)	0.6986	0.5636	0.1155	0.5944	0.5347	0.2644	0.1155	0.5944	0.9182	0.8980	0.0127	0.9281
SegMaR Jia et al. (2022)	0.7796	0.6774	0.0551	0.8691	0.9197	0.8995	0.0152	0.9587	0.9332	0.9179	0.0103	0.9725
FRINet Xie et al. (2023)	0.8539	0.8181	0.0406	0.9269	0.9349	0.9245	0.0111	0.9778	0.9420	0.9350	0.0099	0.9811
HitNet Hu et al. (2023)	0.8121	0.7656	0.0478	0.8953	0.9225	0.9075	0.0120	0.9757	0.9342	0.9235	0.0099	0.9761
PUENet Zhang et al. (2023a)	0.8540	0.8243	0.0323	0.9285	0.9188	0.9078	0.0107	0.9740	0.9335	0.9316	0.0090	0.9810
FSNet Song et al. (2023)	0.8950	0.8709	0.0214	0.9515	0.9423	0.9342	0.0091	0.9823	0.9444	0.9375	0.0088	0.9825
FSEL Sun et al. (2024a)	0.8937	0.8687	0.0209	0.9467	0.9382	0.9309	0.0079	0.9782	0.9390	0.9327	0.0078	0.9773
GLCONetSun et al. (2024b)	0.8831	0.8570	0.0237	0.9410	0.9275	0.9147	0.0103	0.9723	0.9262	0.9117	0.0105	0.9684
PRNetHu et al. (2024)	0.8825	0.8538	0.0258	0.9411	0.9366	0.9314	0.0097	0.9813	0.9394	0.9372	0.0092	0.9817
CamoDiffusionSun et al. (2025)	0.8530	0.8126	0.0416	0.9142	0.9408	0.9309	0.0086	0.9833	0.9429	0.9340	0.0081	0.9836
Ours	**0.8973**	**0.8907**	**0.0237**	**0.9598**	**0.9494**	**0.9505**	**0.0073**	**0.9894**	**0.9530**	**0.9542**	**0.0068**	**0.9896**

Table 6: Compare with state-of-the-art COD models on EVIAC datasets.

To further evaluate robustness under unseen physical environments, Table 6 presents comparisons on EVIAC. PhysCorr consistently outperforms existing methods across fog, forest, and snow scenarios. In particular, under fog conditions where visible contrast degradation and infrared reliability reduction jointly affect target perception, PhysCorr achieves $S_m = 0.8956$ and $F_\beta^w = 0.8921$, indicating stronger robustness against modality degradation. Similar improvements are observed in forest and snow environments, demonstrating that the proposed physical correction mechanism can effectively adapt to different degradation patterns without requiring environment-specific models.

A.3.2 Qualitative Comparisons

In addition to quantitative evaluation, Fig. 6 provides additional qualitative comparisons under conventional and unseen environments. Existing methods often produce incomplete structures, inaccurate boundaries, or false responses when modality reliability changes under different physical conditions.

For conventional environments, although several methods achieve competitive performance, they may still introduce boundary errors or lose fine camouflage details. For example, CGFINet produces relatively accurate predictions under normal conditions but remains sensitive to local ambigu-

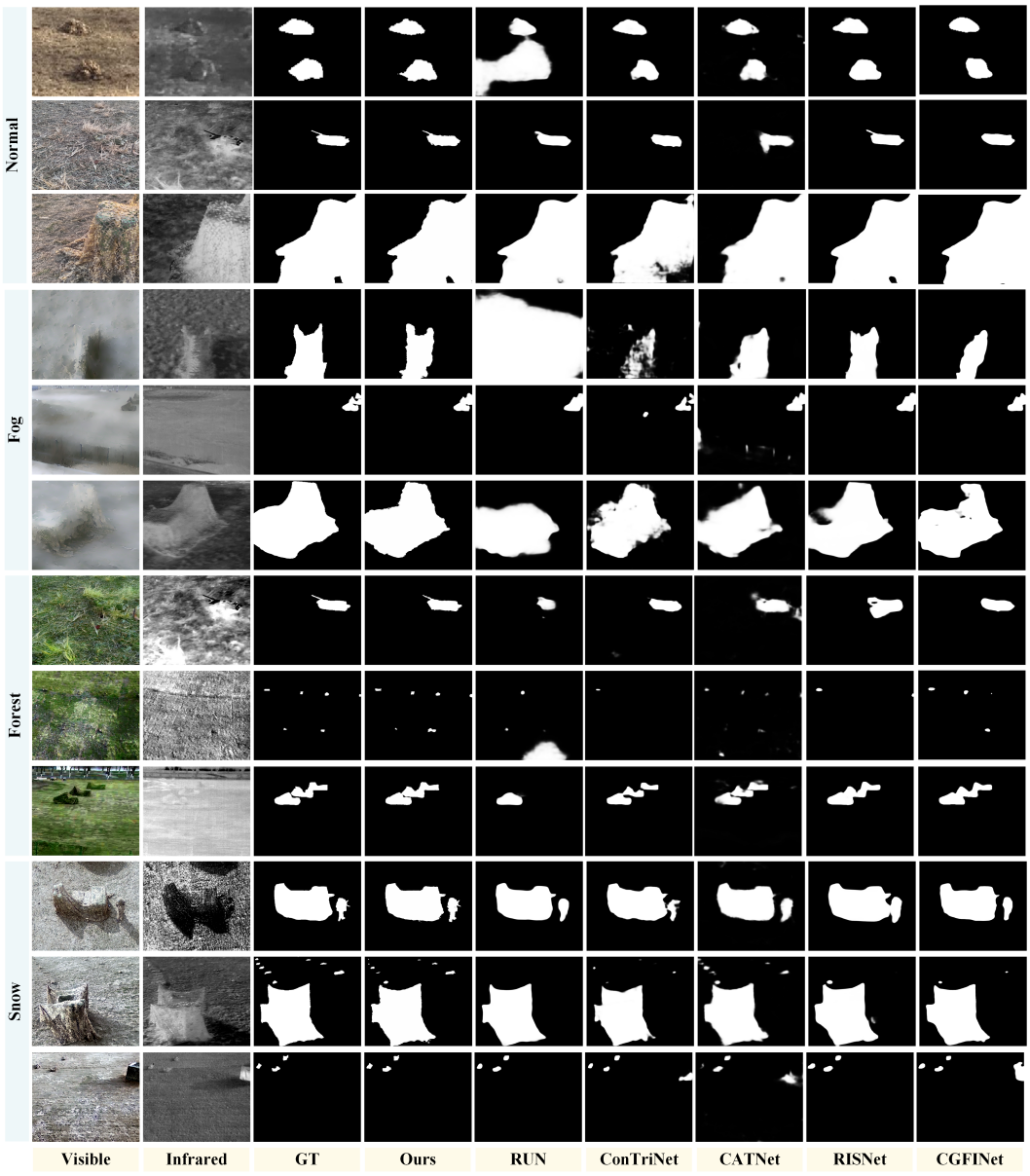

Figure 6: Qualitative comparison on the VIAC and EVIAC datasets. The first three lines show results under conventional environments, while the remaining rows correspond to representative unseen environments, including fog, forest, and snow.

ous regions, indicating that static cross-modal fusion cannot fully model spatially varying modality reliability.

Under fog conditions, RUN tends to lose degraded target structures due to weakened visible cues, while CGFINet may propagate unreliable cross-modal information and generate inaccurate responses around ambiguous regions. In forest scenes, CATNet and RISNet are more likely to confuse complex background textures with camouflage targets, resulting in fragmented predictions or missing regions. Under snow conditions, existing fusion-based methods still suffer from boundary distortion or incomplete segmentation when visible texture is weakened and thermal responses become unstable.

Methods	Params. (M)↓	FLOPs (G)↓
CATNet Sun et al. (2023)	262.74	172.07
XMSNet Wu et al. (2023c)	167.07	74.30
DaCOD Wang et al. (2023)	186.26	128.30
PopNet Wu et al. (2023b)	198.90	228.45
SACNet Wang et al. (2024a)	300.27	143.79
DFINet Wang et al. (2025)	482.40	220.20
Ours	143.01	192.02

Table 7: Comparison of model complexity.

In contrast, PhysCorr produces more complete and accurate segmentation masks across different environments. This improvement is attributed to the reliability-controlled physical correction mechanism. Instead of directly aggregating multimodal features, PhysCorr preserves stable visual evidence and introduces physical correction only in regions where additional guidance is beneficial. Therefore, reliable structures are maintained under conventional environments, while degraded regions are selectively recovered under unseen conditions. These results further demonstrate that physical knowledge serves as adaptive corrective evidence rather than an independent prediction source.

A.4 Complexity and Efficiency Analysis

We further analyze the computational complexity of PhysCorr in terms of model parameters and floating-point operations (FLOPs), as summarized in Table 7. Compared with existing camouflage object segmentation methods, PhysCorr achieves a favorable trade-off between segmentation performance and computational cost. Although some lightweight methods achieve lower FLOPs, they usually rely on simpler feature interaction strategies and cannot provide explicit physical correction under unseen environments. It is worth noting that the vision-language model used for physical prior generation is only involved in offline physical knowledge preparation. During training and inference, PhysCorr directly retrieves the cached physical states without additional vision-language model computation. Therefore, the reported complexity reflects the actual cost of the online segmentation framework. These results demonstrate that PhysCorr introduces physical reasoning capability while maintaining a practical computational cost for visible–infrared camouflage segmentation.

A.5 Theoretical Interpretation

The proposed PhysCorr can be interpreted as a reliability-guided residual correction process. Let Z^* denote the ideal prediction logit under the unknown target environment. The visual branch provides a stable estimation Z_v, while the physical branch estimates an environment-conditioned correction term. Instead of directly replacing the visual prediction, PhysCorr models the final prediction as:

$$Z = Z_v + \rho q S \odot \Delta, \tag{16}$$

where q and S represent the correction confidence and spatial support, respectively. The correction term is constrained by the reliability controller, which prevents uncertain physical information from being uniformly propagated.

When the visual evidence is reliable, the controller decreases the correction magnitude:

$$qS \to 0, \tag{17}$$

and the prediction approaches the stable visual estimation:

$$Z \to Z_v. \tag{18}$$

In contrast, when modality degradation causes the visual evidence to become unreliable, the controller increases the contribution of physical guidance:

$$qS \to 1, \tag{19}$$

allowing the physical branch to provide complementary correction. Therefore, the residual formulation naturally preserves reliable structures while recovering degraded target regions.

Furthermore, the counterfactual invariant learning constrains the physical representation under different environmental states. Given observed and counterfactual physical states s and s^{cf}, the objec-

tive encourages:

$$d(z_r(s), z_r(s^{cf})) \to 0, \tag{20}$$

which reduces the dependence of physical representations on environment- specific degradation factors and improves generalization to unseen conditions.

A.6 Discussion

PhysCorr provides a different perspective for visible–infrared camouflage object segmentation under complex environments. Instead of treating multimodal features as equally reliable information, PhysCorr explicitly models the physical factors that cause modality degradation and uses physical knowledge as adaptive corrective evidence. This design enables the model to distinguish stable target-related information from environment-induced variations, thereby avoiding the propagation of unreliable modality cues during feature fusion. The effectiveness of PhysCorr lies in the combination of structured physical reasoning and reliability-controlled correction. The physical state provides high-level guidance about environment conditions and modality reliability, while the correction controller determines where and how physical information should influence the visual prediction. Therefore, the proposed framework can preserve reliable structures under normal conditions and recover degraded target regions under unseen environments. These results suggest that incorporating interpretable physical knowledge into multimodal perception is a promising direction for improving robustness in challenging real-world scenarios.

A.7 Potential Applications

The proposed PhysCorr framework has potential applications in visible–infrared perception systems operating under complex and uncertain environments. By explicitly modeling environment-induced modality degradation and adaptively correcting unreliable information, PhysCorr can provide more robust target perception compared with conventional multimodal fusion strategies.

In practical scenarios, such as autonomous aerial surveillance, camouflage target detection, border monitoring, and search-and-rescue missions, visible and infrared sensors often suffer from different environmental disturbances, including fog, snow, illumination variation, and background interference. The proposed physical-state modeling and reliability-controlled correction mechanism can help maintain stable target localization when one modality becomes unreliable. Therefore, PhysCorr provides a potential solution for robust visible–infrared perception in safety-critical applications where reliable target segmentation is required under diverse physical conditions.