GOLD: GENERALIZED PROPOSAL DENOISING FOR DOMAIN-GENERALIZED OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Domain-generalized object detection (DGOD) aims to train a detector solely on labeled source domain data that generalizes directly to unseen target domains without accessing target data or performing test-time adaptation. Despite encouraging progress, existing approaches largely optimize encoded features or final predictions within deterministic detection frameworks and thus lack an explicit mechanism to iteratively recover object locations from unreliable spatial hypotheses. To this end, we propose **G**eneralized Pr**O**posa**L** **D**enoising (**GOLD**), a diffusion-based DGOD framework that reformulates cross-domain localization as a semantics-preserving denoising process that progressively refines noisy proposals under unreliable visual evidence. First, a Spatial Calibration Adapter (SCAdapter) equips frozen DINOv3 features with localization-sensitive spatial structure for iterative proposal refinement. Second, Spatially Heterogeneous Evidence Perturbing (SHEP) decomposes visual evidence into stochastic foreground regions, localization-critical contexts, and distant backgrounds, as well as applies moderate, strong, and weak residual perturbations, respectively, according to their roles in proposal denoising. Third, Prototype-Anchored Denoising (PAD) constructs multi-scale category prototypes and anchors matched proposal representations at every refinement stage, preserving category semantics throughout proposal denoising. Extensive experiments on multiple cross-domain benchmarks demonstrate that GOLD consistently outperforms previous DGOD methods. *Our code is released in supplementary materials.*

1 INTRODUCTION

Object detectors often achieve high accuracy within their training domains, yet their performance can degrade substantially when test images exhibit variations in weather, illumination, or imaging devices. This limitation severely restricts their deployment in open and dynamically changing environments. Therefore, Domain-Generalized Object Detection (DGOD) aims to train a detector solely on labeled source domain data and directly generalize it to unseen domains without accessing any target domain data Lin et al. (2021).

Existing DGOD methods can be broadly categorized into three groups. The first group comprises domain-diversification methods, which enlarge the source domain distribution through image- or feature-level perturbations, synthesized styles, or diffusion-based feature augmentation Danish et al. (2024); Zhang et al. (2025); Jiang et al. (2025). The second group learns causal representations by disentangling domain-specific factors with causal objectives, suppressing non-causal visual cues, and aligning representations across diversified domains Wu & Deng (2022); Xu et al. (2023); Liu et al. (2024). The third group leverages Vision Foundation Models (VFMs), whose semantic priors and generalization capabilities provide detectors with more transferable representation supervision Vidit et al. (2023); Xu et al. (2025); Zhang et al. (2026c). Despite encouraging progress, these methods primarily optimize encoded features or final predictions within conventional discriminative detection frameworks, which lack an explicit object-locating mechanism for iteratively recovering object locations from unreliable spatial hypotheses.

Fortunately, DiffusionDet Chen et al. (2023) offers a different perspective for addressing this limitation. It formulates object detection as a generative denoising process from random noisy boxes to ground-truth boxes and progressively refines proposal locations and scales through a multi-stage

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.

detection head. However, DiffusionDet models uncertainty primarily by adding Gaussian noise to proposal coordinates, while assuming that the visual features guiding iterative refinement remain reliable. This assumption may not hold under unseen domain shifts, where visual evidence can be unevenly degraded across spatial regions. Existing domain-diversification methods typically overlook such spatial heterogeneity, potentially under-perturbing localization-sensitive contextual evidence while over-corrupting the category-discriminative foreground evidence required for detection.

Motivated by this observation, we propose **G**eneralized Pr**O**posa**L** **D**enoising (**GOLD**), a diffusion-based detection framework that enables reliable iterative localization under different domain shifts. Concretely, the Spatial Calibration Adapter (SCAdapter) injects localization-sensitive spatial structure into frozen DINOv3 features before multi-scale fusion. Then, Spatially Heterogeneous Evidence Perturbing (SHEP) decomposes visual evidence according to its role in proposal denoising and applies the strongest perturbation to localization-critical contexts, a moderate perturbation to stochastic foreground regions, and the weakest perturbation to distant backgrounds. This design concentrates stronger training perturbations on evidence used for proposal refinement while preserving category-discriminative cues.

Finally, Prototype-Anchored Denoising (PAD) constructs robust multi-scale category prototypes from raw DINOv3 features and anchors matched proposal representations at every refinement stage. These raw semantic references stabilize category semantics as noisy boxes are progressively recovered. Extensive experiments on multiple DGOD benchmarks demonstrate that GOLD consistently improves cross-domain detection performance and achieves state-of-the-art results.

Our main contributions are summarized as follows:

- We propose GOLD, a diffusion-based detection framework that reformulates domain-generalized object detection as a semantics-preserving denoising process that progressively refines noisy proposals under unreliable visual evidence.
- GOLD integrates three complementary components: SCAdapter equips frozen DINOv3 features with localization-sensitive spatial structure, SHEP diversifies spatial evidence according to its role in proposal refinement, and PAD anchors proposal representations to multi-scale category prototypes throughout denoising.
- Extensive experiments demonstrate that GOLD consistently outperforms prior methods and achieves SOTA performance across diverse unseen domains.

2 Related Works

Domain-generalized object detection (DGOD) trains a detector on labeled source-domain data and evaluates it directly on previously unseen domains. Existing approaches mainly pursue domain diversification, causal invariance, and foundation-model transfer. For domain diversification, OA-DG (Lee et al., 2024) synthesizes object-aware domains, while DivAlign (Danish et al., 2024) combines source diversification with prediction-level alignment for both classification and localization. For causal invariance, cyclic-disentangled self-distillation (Wu & Deng, 2022) learns domain-invariant object representations, whereas MAD (Xu et al., 2023) mines non-causal factors through multi-view adversarial learning to reduce reliance on domain-specific cues. For foundation-model transfer, CLIP the Gap (Vidit et al., 2023) introduces language-defined domain concepts into detector features, and GDD (He et al., 2025b) transfers multi-step representations from a pretrained diffusion model through feature- and object-level alignment. Cauvis (Li et al., 2025) combines causal visual prompts with a dual-branch adapter to suppress spurious visual evidence, while Bridge (Hong et al., 2026) learns compact basis representations for front-door adjustment and integrates causal intervention with discriminative and generative VFMs. Prototype learning complements these directions by preserving category structure when object appearances change across domains. UFR (Liu et al., 2024) uses causal attention and prototype constraints to mitigate scene- and object-attribute confounders. Style Evolving along Chain-of-Thought (Zhang et al., 2025) combines source-style expansion with class-specific prototype clustering, while MR-DCoT (Zhang et al., 2026e) anchors category prototypes on a semantic manifold and rectifies off-manifold features toward class-conditional neighborhoods.

Taken together, these methods improve the diversity of source observations or the transferability of learned representations, but give less attention to the reliability of visual evidence during successive

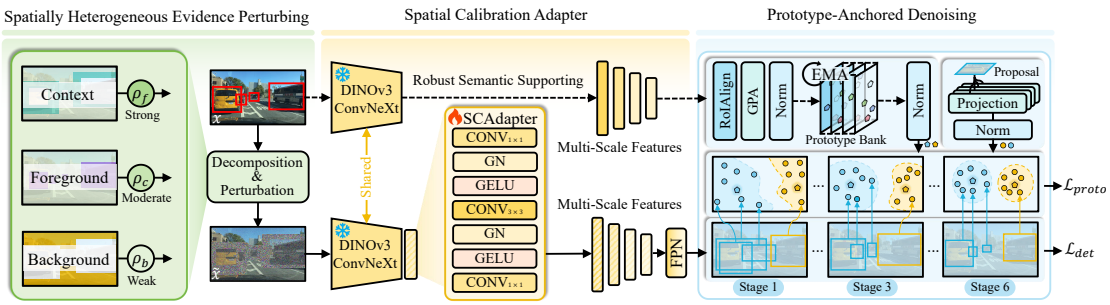

Figure 1: Overview of **GOLD**, which consists of three modules: (1) the Spatial Calibration Adapter (SCAdapter) injects localization-sensitive spatial structure into frozen DINOv3 features; (2) Spatially Heterogeneous Evidence Perturbing (SHEP) diversifies spatial evidence according to its role in proposal denoising; and (3) Prototype-Anchored Denoising (PAD) preserves category semantics throughout the proposal denoising stages.

localization updates. Under an unseen shift, a detector may retain cues for recognizing an object while misreading the boundaries and context needed to refine its box. GOLD therefore approaches DGOD as a problem of maintaining dependable visual guidance throughout iterative proposal denoising, so category cues remain stable as uncertain object locations are progressively corrected.

3 Preliminaries

3.1 Task Formulation

Domain-generalized object detection (DGOD) trains a detector using only a labeled source domain and directly evaluates it on previously unseen target domains. Given a source dataset $\mathcal{D}_s = \{(x_i, \mathcal{Y}_i)\}_{i=1}^{N_s}$, where $\mathcal{Y}_i$ contains the bounding boxes and category labels of image x_i, the objective is to learn a detector that remains effective under unknown domain shifts.

3.2 DiffusionDet

DiffusionDet (Chen et al., 2023) formulates object detection as a box-denoising process. During training, ground-truth boxes are corrupted by Gaussian noise sampled according to a diffusion schedule. Conditioned on the image features, noisy boxes, and timestep embedding, the detector predicts category scores and reconstructs the clean object boxes. DiffusionDet adopts optimal-transport assignment to match predictions with ground-truth objects. The detection objective is applied to all six refinement stages:

$$\mathcal{L}_{\text{det}} = \sum_{s=1}^{S} \left(\lambda_{\text{cls}} \mathcal{L}_{\text{focal}}^{(s)} + \lambda_{\ell_1} \mathcal{L}_{\ell_1}^{(s)} + \lambda_{\text{giou}} \mathcal{L}_{\text{GIoU}}^{(s)} \right), \tag{1}$$

where $S = 6$, and the loss weights are set to $\lambda_{\text{cls}} = 2.0$, $\lambda_{\ell_1} = 5.0$, and $\lambda_{\text{giou}} = 2.0$.

4 Method

4.1 Overview

As illustrated in Fig. 1, GOLD comprises three complementary modules. Firstly, SCAdapter calibrates the frozen foundation model features with local spatial structure before multi-scale fusion and iterative proposal denoising. Then, SHEP decomposes each source image into foreground, context, and background regions, and applies moderate, strong, and weak coefficients, respectively, to a shared perturbation residual. Finally, PAD constructs clean multi-scale category prototypes from ground-truth RoI features of the raw DINOv3 hierarchy and anchors matched proposal representations at all six refinement stages. Together, these modules expose proposal denoising to spatially het-

erogeneous evidence perturbations while preserving semantic-consistent representations throughout localization recovery.

4.2 Spatial Calibration Adapter

DINOv3 Siméoni et al. (2025) provides transferable dense representations learned from large-scale self-supervised pretraining. We employ a frozen DINOv3 ConvNeXt encoder to extract four feature levels $\{R_l\}_{l=2}^{5}$ with strides $\{4, 8, 16, 32\}$. Freezing the encoder preserves its pretrained semantic space and prevents it from overfitting to the appearance statistics of the single source domain. Nevertheless, these features are not explicitly optimized for the spatial localization required by iterative proposal refinement. To bridge this semantic–localization mismatch, we introduce a trainable Spatial Calibration Adapter (SCAdapter) at each feature level.

Let $\phi(\cdot) = \delta(\mathrm{GN}(\cdot))$ denote the normalization and activation operation. The calibration operation is formulated as:

$$\begin{aligned} Z_l &= \phi\big(\mathcal{P}_l^{\mathrm{in}}(R_l)\big)\,, \\ \widetilde{R}_l &= \mathcal{P}_l^{\mathrm{out}}(\phi(\mathrm{DWConv}_{3\times3}(Z_l)))\,, \end{aligned} \tag{2}$$

where Z_l denotes the bottleneck feature, and $\mathcal{P}_l^{\mathrm{in}}$ and $\mathcal{P}_l^{\mathrm{out}}$ are 1×1 projections for channel compression and restoration, respectively. The depthwise convolution $\mathrm{DWConv}_{3\times3}$ injects local spatial structure into the frozen representation with limited additional parameters, while GN and δ denote Group Normalization and GELU, respectively. Within the detection branch, an FPN fuses the spatially calibrated multi-scale features $\{\widetilde{R}_l\}_{l=2}^{5}$, and the resulting feature pyramid is fed into the detection head for iterative proposal refinement.

4.3 Spatially Heterogeneous Evidence Perturbing

At each refinement stage, DiffusionDet estimates box updates from proposal-aligned features. However, foreground cues support category recognition, while boundaries and foreground–background transitions guide box refinement. Uniform perturbation cannot control these cues separately, as weak perturbations barely affect localization cues, while strong ones may disrupt category semantics. Therefore, we propose Spatially Heterogeneous Evidence Perturbing (SHEP), which comprises Spatially Heterogeneous Evidence Decomposition and Denoising-Oriented Evidence Perturbation. The former partitions each image into three functional evidence regions, while the latter modulates a shared perturbation residual according to each region's role in proposal denoising.

Spatially Heterogeneous Evidence Decomposition. Let Ω denote the image domain and let the ground-truth box of object i be $b_i = (\mathbf{c}_i, \mathbf{d}_i)$ in center-size form, where $\mathbf{c}_i = (c_i^x, c_i^y)$ is the box center and $\mathbf{d}_i = (w_i, h_i)$ contains its width and height. We sample a shared scale factor $s_i \sim \mathcal{U}(s_-, s_+)$ and independent center offsets $\epsilon_i^x, \epsilon_i^y \overset{\text{i.i.d.}}{\sim} \mathcal{U}(-\eta, \eta)$. The perturbed center and size are:

$$\widehat{\mathbf{c}}_i = (c_i^x + \epsilon_i^x w_i,\ c_i^y + \epsilon_i^y h_i)\,, \qquad \widehat{\mathbf{d}}_i = (\max\{s_i w_i, d_{\min}\},\ \max\{s_i h_i, d_{\min}\})\,. \tag{3}$$

These parameters form an unclipped stochastic foreground-support box $\bar{b}_i$, which is subsequently clipped to the image domain:

$$\bar{b}_i = \left(\widehat{\mathbf{c}}_i, \widehat{\mathbf{d}}_i\right), \qquad \widehat{b}_i = \mathrm{Clip}_{\Omega}(\bar{b}_i). \tag{4}$$

Here, $\mathrm{Clip}_{\Omega}(\cdot)$ clips a box to the image boundaries. The condition $s_- < 1 < s_+$ allows $\widehat{b}_i$ to undercover or over-cover the ground-truth object, while the center offsets simulate translation errors. The minimum size $d_{\min}$ prevents the foreground supports of small objects from vanishing. We therefore treat $\widehat{b}_i$ as a stochastic, box-derived foreground support.

Next, we construct a localization context enclosing both b_i and $\widehat{b}_i$. Let $\mathcal{H}(b_i, \widehat{b}_i)$ denote their smallest axis-aligned enclosing rectangle. We independently sample $\lambda_i^x, \lambda_i^y \overset{\text{i.i.d.}}{\sim} \mathcal{U}(\lambda_-, \lambda_+)$ and define the horizontal and vertical expansion margins as:

$$\mathbf{m}_i = (m_i^x, m_i^y) = (\max\{\lambda_i^x w_i, m_{\min}\},\ \max\{\lambda_i^y h_i, m_{\min}\})\,. \tag{5}$$

The outer context box is

$$b_i^{\mathrm{ctx}} = \mathrm{Clip}_{\Omega}\left(\mathrm{Expand}\left(\mathcal{H}(b_i, \widehat{b}_i), \mathbf{m}_i\right)\right). \tag{6}$$

Here, $\mathrm{Expand}(\cdot, \mathbf{m}_i)$ extends the enclosing rectangle by m_i^x on both horizontal sides and by m_i^y on both vertical sides. Therefore, b_i^{ctx} covers the discrepancy between the ground-truth and sampled boxes, together with the surrounding boundary evidence used for box correction. It defines the outer context extent; the actual context region is obtained after excluding all foreground-support pixels.

Let $\mathcal{P}(b)$ denote the rasterized support of box b, and let $\mathbb{I}[\cdot]$ be the indicator function. We aggregate the foreground supports and context extents as $\mathcal{F} = \bigcup_i \mathcal{P}(\widehat{b}_i)$ and $\mathcal{E} = \bigcup_i \mathcal{P}(b_i^{\mathrm{ctx}})$, respectively. The three region masks are then defined as:

$$M_{\mathrm{fg}}(p) = \mathbb{I}[p \in \mathcal{F}], \quad M_{\mathrm{ctx}}(p) = \mathbb{I}[p \in \mathcal{E} \setminus \mathcal{F}], \quad M_{\mathrm{bg}}(p) = \mathbb{I}[p \in \Omega \setminus \mathcal{E}]. \tag{7}$$

Because every sampled foreground support is contained in its corresponding outer context box, $\mathcal{F} \subseteq \mathcal{E}$. The three masks are therefore mutually exclusive and cover Ω. Moreover, removing the global foreground union $\mathcal{F}$ from $\mathcal{E}$ gives foreground regions priority, preventing the foreground support of one instance from being treated as the context of a neighboring instance in crowded scenes. These masks provide heterogeneous evidence supports for category semantics, localization, and scene context, respectively. Representative examples of the resulting three-region decomposition are shown in Figure 2.

Denoising-Oriented Evidence Perturbation. We sample a single photometric transformation $\mathcal{T}$ composed of color jittering, Gaussian blur, and Gaussian noise, and define its residual as $\Delta x = \mathcal{T}(x) - x$. We spatially modulate this shared residual using the three masks:

$$\widetilde{x} = x + (\rho_{\mathrm{fg}} M_{\mathrm{fg}} + \rho_{\mathrm{ctx}} M_{\mathrm{ctx}} + \rho_{\mathrm{bg}} M_{\mathrm{bg}}) \odot \Delta x. \tag{8}$$

We set $0 \le \rho_{\mathrm{bg}} < \rho_{\mathrm{fg}} < \rho_{\mathrm{ctx}} \le 1$. The localization context receives the strongest perturbation to challenge boundary-dependent box refinement; the stochastic foreground region receives a moderate perturbation to diversify object appearance while retaining category-discriminative evidence; and the distant background receives the weakest perturbation to preserve coherent scene statistics. Because all transformations are photometric, $\widetilde{x}$ remains geometrically aligned with the original annotations. The spatially perturbed view $\widetilde{x}$ is used for detector training, whereas the clean view x supplies ground-truth RoI descriptors for Multi-Scale Prototype Construction.

4.4 Prototype-Anchored Denoising

Although SHEP improves robustness to unreliable visual evidence, it may also shift intermediate proposal representations toward perturbed appearance cues. We therefore introduce Prototype-Anchored Denoising (PAD), which constructs multi-scale category prototypes as stable semantic references and anchors matched proposal representations at every refinement stage.

Multi-Scale Prototype Construction. Given a clean source image, the frozen DINOv3 encoder produces four raw feature maps $\{R_l^{\mathrm{cln}}\}_{l=2}^{5}$. For every ground-truth object (b_i, y_i), we independently apply RoIAlign to all four levels rather than assigning it to a single FPN level:

$$z_{i,l} = \mathrm{Norm}\left(\mathrm{GAP}\left[\mathrm{RoIAlign}\left(R_l^{\mathrm{cln}}, b_i\right)\right]\right), \qquad l \in \{2,3,4,5\}, \tag{9}$$

where GAP denotes global average pooling and Norm denotes ℓ_2 normalization. Let $\mathcal{I}_c = \{i \mid y_i = c\}$ denote the ground-truth instances of category c in the current mini-batch. At each feature level l, their clean instance descriptors are aggregated into a batch class center:

$$\bar{z}_{c,l} = \mathrm{Norm}\left(\frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} z_{i,l}\right). \tag{10}$$

We maintain a multi-scale prototype bank $\mathcal{P} = \{p_{c,l}\}$ using these centers. For every category present in the current mini-batch, its level-l prototype is initialized or updated as

$$p_{c,l} \leftarrow \begin{cases} \bar{z}_{c,l}, & \text{if } p_{c,l} \text{ is uninitialized,} \\ \mathrm{Norm}(\mu p_{c,l} + (1-\mu)\bar{z}_{c,l}), & \text{otherwise,} \end{cases} \tag{11}$$

where μ is the EMA momentum. Prototypes of categories absent from the current mini-batch remain unchanged. The bank is updated without gradients using only clean ground-truth RoI descriptors.

Prototype-Guided Proposal Denoising. Within PAD, let h_q^s denote the representation of proposal q at detection stage s, extracted after proposal interaction and feature transformation but before the classification and regression branches. At each stage, we reuse the original DiffusionDet assignment to identify positive proposal–ground-truth pairs. Unmatched background proposals are excluded from prototype-guided supervision.

Because the four prototype levels have different channel dimensions, scale-specific projectors map each matched proposal representation into the corresponding DINOv3 feature space:

$$\bar{h}_{q,l}^s = \mathrm{Norm}\big(g_l(h_q^s)\big), \qquad a_{q,c}^{s,l} = \frac{\left(\bar{h}_{q,l}^s\right)^\top \mathrm{sg}(p_{c,l})}{\tau}, \tag{12}$$

where g_l is a lightweight projection layer, τ is the temperature, and sg denotes stop-gradient. For the matched proposal index set $\mathcal{M}_s$, let y_q denote the category assigned to proposal q and $\mathcal{V}_l$ the classes with initialized level-l prototypes. With $S = 6$, the multi-stage prototype-guided proposal-denoising objective is:

$$\mathcal{L}_{\mathrm{proto}} = -\sum_{s=1}^{S}\omega_s \sum_{l=2}^{5}\beta_l \frac{1}{|\mathcal{M}_s|}\sum_{q\in\mathcal{M}_s}\log\frac{\exp\left(a_{q,y_q}^{s,l}\right)}{\sum_{c\in\mathcal{V}_l}\exp\left(a_{q,c}^{s,l}\right)}, \tag{13}$$

where ω_s and β_l control the contributions of different stages and feature levels. Later stages receive larger weights because their proposal assignments are more reliable. Since the prototypes are detached while proposal representations retain gradients, this objective anchors proposal representations to their corresponding clean category prototypes while separating them from prototypes of other categories. The semantically regularized representations subsequently produce the next box updates, thereby coupling semantic preservation with iterative localization.

4.5 Overall Optimization Objective

GOLD is jointly optimized using the original DiffusionDet detection objective and the prototype-guided proposal-denoising objective:

$$\mathcal{L}_{\mathrm{GOLD}} = \mathcal{L}_{\mathrm{det}} + \lambda_{\mathrm{proto}}\mathcal{L}_{\mathrm{proto}}, \tag{14}$$

where $\mathcal{L}_{\mathrm{det}}$ is evaluated on the spatially heterogeneous perturbed view $\widetilde{x}$, and λ_{proto} controls the contribution of Prototype-Guided Proposal Denoising. During inference, SHEP and the PAD-specific supervision branch are removed, while the frozen DINOv3 backbone, SCAdapter, and detection head are retained. We provide the detailed pseudocode of our training process in Appendix C.

5 Experiment

5.1 Datasets and Evaluation Metric

The Diverse Weather Dataset Wu & Deng (2022) covers diverse adverse weather and illumination conditions. Models are trained on daytime-sunny (DS) and evaluated on daytime-foggy (DF), dusk-rainy (DR), night-rainy (NR), and night-sunny (NS). We also evaluate Cityscapes Cordts et al. (2016)→BDD100K Yu et al. (2020) and Cityscapes→Foggy Cityscapes Sakaridis et al. (2018), using the BDD100K daytime-clear split and Foggy Cityscapes with the most challenging attenuation coefficient split 0.02 as unseen targets, respectively. We use the mean Average Precision (mAP) with a threshold of 0.5 to evaluate the detection performance.

Table 1: DG results (%) on Diverse Weather Datasets. All entries are mAP. Avg. is the arithmetic mean over the four unseen target domains.

Method	Venue	DF	DR	NR	NS	Avg.
S-DGOD Wu & Deng (2022)	CVPR'22	33.5	28.2	16.6	36.6	28.7
CLIP the Gap Vidit et al. (2023)	CVPR'23	38.5	32.3	18.7	36.9	31.6
SRCD Rao et al. (2024)	TNNLS'24	35.9	28.8	17.0	36.7	29.6
PDOC Li et al. (2024)	CVPR'24	39.1	33.7	19.2	38.5	32.6
UFR Liu et al. (2024)	CVPR'24	39.6	33.2	19.2	40.8	33.2
G-NAS Wu et al. (2024)	AAAI'24	36.4	35.1	17.4	45.0	33.5
DivAlign Danish et al. (2024)	CVPR'24	37.2	38.1	24.1	42.5	35.5
FWCL Guo et al. (2025)	TMM'25	32.3	32.6	18.9	37.5	30.3
GDD He et al. (2025b)	CVPR'25	43.3	42.5	27.8	47.0	40.2
Boost He et al. (2025a)	ICCV'25	45.9	44.3	28.2	48.8	41.8
SE-COT Zhang et al. (2025)	CVPR'25	44.9	49.5	33.7	52.7	45.2
DoSL Zhang et al. (2026b)	CVPR'26	41.2	40.6	24.8	44.0	37.7
CD-FKD Lee et al. (2026)	ICRA'26	40.2	42.3	23.4	47.3	38.3
UICL Zhang et al. (2026a)	AAAI'26	40.1	42.2	27.4	45.2	38.7
LTFE Zhang et al. (2026d)	AAAI'26	53.1	46.5	33.4	46.4	44.9
VFM^{4}SDG Zhang et al. (2026c)	arXiv'26	47.4	50.7	29.0	56.1	45.8
MR-DCoT Zhang et al. (2026e)	TPAMI'26	48.3	53.2	36.4	54.6	48.1
Bridge Hong et al. (2026)	CVPR'26	48.1	53.8	46.7	54.6	50.8
DiffusionDet Chen et al. (2023)	ICCV'23	42.3	49.2	28.7	51.2	42.9
GOLD (Ours)		**55.6**	**62.3**	**48.8**	**60.5**	**56.8**

Table 2: DG Results (%) on Daytime-Foggy.

Method	Venue	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
S-DGOD	CVPR'22	32.9	28.0	48.8	29.8	32.5	38.2	24.1	33.5
CLIP the Gap	CVPR'23	36.1	34.3	58.0	33.1	39.0	43.9	25.1	38.5
PDOC	CVPR'24	36.1	34.5	58.4	33.3	40.5	44.2	26.2	39.1
UFR	CVPR'24	36.9	35.8	61.7	33.7	39.5	42.2	27.5	39.6
SE-COT	CVPR'25	42.1	38.3	62.8	40.8	47.5	48.0	35.4	45.0
CD-FKD	ICRA'26	36.6	33.2	62.6	34.6	42.9	44.2	27.5	40.2
MR-DCoT	TPAMI'26	44.3	40.4	67.3	38.0	50.3	**59.1**	38.5	48.3
DiffusionDet	ICCV'23	42.9	31.6	59.7	39.0	43.6	44.9	34.8	42.3
GOLD (Ours)		**56.0**	**44.8**	**73.0**	**52.0**	**56.5**	58.3	**48.3**	**55.6**

Table 3: DG Results (%) on Dusk-Rainy.

Method	Venue	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
S-DGOD	CVPR'22	37.1	19.6	50.9	13.4	19.7	16.3	40.7	28.2
CLIP the Gap	CVPR'23	37.8	22.8	60.7	16.8	26.8	18.7	42.4	32.3
UFR	CVPR'24	37.1	21.8	67.9	16.4	27.4	17.9	43.9	33.2
PDOC	CVPR'24	39.4	25.2	60.9	20.4	29.9	16.5	43.9	33.7
SE-COT	CVPR'25	57.3	40.2	76.8	28.4	50.0	32.7	60.8	49.5
CD-FKD	ICRA'26	50.3	30.9	75.7	19.6	40.4	26.4	52.9	42.3
MR-DCoT	TPAMI'26	61.6	41.0	79.1	31.3	57.3	37.4	64.8	53.2
DiffusionDet	ICCV'23	52.4	42.5	71.2	35.7	51.3	33.7	57.7	49.2
GOLD (Ours)		**66.0**	**55.9**	**84.3**	**49.5**	**64.3**	**45.6**	**70.5**	**62.3**

Table 4: DG Results (%) on Night-Rainy.

Method	Venue	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
S-DGOD	CVPR'22	24.4	11.6	29.5	9.8	10.5	11.4	19.2	16.6
CLIP the Gap	CVPR'23	28.6	12.1	36.1	9.2	12.3	9.6	22.9	18.7
PDOC	CVPR'24	25.6	12.1	35.8	10.1	14.2	12.9	22.9	19.2
UFR	CVPR'24	29.9	11.8	36.1	9.4	13.1	10.5	23.3	19.2
SE-COT	CVPR'25	49.3	19.1	56.7	11.1	31.3	20.9	47.8	33.7
CD-FKD	ICRA'26	34.3	14.2	49.6	2.1	17.2	15.6	30.6	23.4
MR-DCoT	TPAMI'26	50.3	21.4	60.3	17.6	33.3	20.1	52.0	36.4
DiffusionDet	ICCV'23	44.4	12.0	47.7	12.7	28.8	14.1	41.4	28.7
GOLD (Ours)		**64.1**	**31.4**	**67.3**	**35.5**	**48.7**	**33.6**	**61.0**	**48.8**

Table 5: DG Results (%) on Night-Sunny.

Method	Venue	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
S-DGOD	CVPR'22	40.6	35.1	50.7	19.7	34.7	32.1	43.4	36.6
CLIP the Gap	CVPR'23	37.7	34.3	58.0	19.2	37.6	28.5	42.9	36.9
PDOC	CVPR'24	40.9	35.0	59.0	21.3	40.4	29.9	42.9	38.5
UFR	CVPR'24	43.6	38.1	66.1	14.7	49.1	26.4	47.5	40.8
SE-COT	CVPR'25	55.5	49.7	71.0	32.9	58.8	43.7	57.8	52.8
CD-FKD	ICRA'26	47.8	40.1	71.8	27.2	55.2	38.8	50.1	47.3
MR-DCoT	TPAMI'26	50.3	53.2	**80.1**	33.5	63.2	41.2	60.8	54.6
DiffusionDet	ICCV'23	54.8	48.1	67.5	32.4	57.9	43.1	54.8	51.2
GOLD (Ours)		**63.7**	**57.9**	76.4	**41.8**	**66.7**	**53.0**	**63.8**	**60.5**

Table 6: DG Results (%) on BDD100K.

Method	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
CDSD Wu & Deng (2022)	20.5	22.9	33.8	14.7	18.5	23.6	18.2	21.7
SHADE Zhao et al. (2022)	19.0	25.1	36.8	18.4	24.1	24.9	19.8	24.0
SRCD Rao et al. (2024)	21.5	24.8	38.7	19.0	25.7	28.4	23.1	25.9
GDD He et al. (2025b)	31.0	38.9	71.5	37.6	61.5	47.0	38.5	46.6
Boost He et al. (2025a)	41.7	41.2	72.7	37.2	62.8	48.7	40.5	49.3
Bridge Hong et al. (2026)	54.2	**49.5**	74.3	55.0	66.5	**58.2**	**54.5**	58.9
DiffusionDet Chen et al. (2023)	33.7	28.5	67.9	25.6	50.1	39.1	28.9	39.1
GOLD (Ours)	**58.3**	43.6	**79.1**	**57.5**	**69.9**	56.7	51.2	**59.5**

Table 7: DG Results (%) on FoggyCityscapes.

Method	Bus	Bike	Car	Motor	Psn.	Rider	Train	Truck	mAP
DIDN Lin et al. (2021)	35.7	33.1	49.3	24.8	31.8	38.4	26.5	27.7	33.4
MAD ?	44.0	40.1	45.0	30.3	34.2	47.4	42.4	25.6	38.6
GDD He et al. (2025b)	53.8	54.2	67.5	45.6	52.1	60.8	53.9	32.4	52.5
Boost He et al. (2025a)	55.3	**62.7**	68.2	45.5	52.9	62.0	48.4	37.4	54.1
UICL Zhang et al. (2026a)	53.1	61.1	69.7	46.3	58.2	62.0	40.5	34.5	53.2
Bridge Hong et al. (2026)	72.1	58.6	73.2	51.7	58.5	64.6	65.4	48.6	61.6
DiffusionDet Chen et al. (2023)	52.4	42.7	57.4	36.2	45.3	51.3	38.3	38.4	45.2
GOLD (Ours)	**74.1**	60.0	**74.7**	**55.4**	**65.1**	**68.7**	**73.4**	**53.2**	**65.6**

5.2 Implementation Details

We implement GOLD based on DiffusionDet (Chen et al., 2023) using a frozen DINOv3 ConvNeXt-L backbone (Siméoni et al., 2025). Concretely, SCAdapter employs bottleneck dimensions of $(64, 128, 256, 512)$ for the four backbone levels. For SHEP, we set $(s^-, s^+) = (0.5, 1.5)$, $\eta = 0.4$, and $(\lambda^-, \lambda^+) = (0.2, 0.7)$. For PAD, the prototype bank is updated with an EMA momentum μ of 0.95. temperature $\tau = 0.07$, loss weight $\lambda_{\text{proto}} = 0.2$, stage weights $\boldsymbol{\omega} = (0.05, 0.05, 0.10, 0.20, 0.25, 0.40)$, and scale weights $\boldsymbol{\beta} = (0.25, 0.25, 0.25, 0.25)$. We train the model on source domain data for 450,000 iterations on one NVIDIA GeForce RTX 4090 GPU, using AdamW with a batch size of 4 and an initial learning rate of 2.5×10^{-5}. We provide detailed parameter settings in Appendix B.

5.3 Comparison with State-of-the-Arts

Diverse Weather. Table 1 compares GOLD with prior domain-diversification, causal-invariance, and VFM-based approaches under four weather and illumination shifts. GOLD achieves an average mAP of 56.8%, surpassing the previous best result of 50.8% mAP by 6.0% mAP. These results demonstrate that our method improves cross-domain generalization by iteratively denoising proposals under degraded visual conditions. Specifically, Table 2 reports the daytime-foggy results, where GOLD achieves 55.6% mAP, surpassing the strongest SOTA by 7.3% mAP and leading in six of seven categories. On dusk-rainy and night-rainy conditions, GOLD obtains 62.3% and 48.8% mAP in Tables 3 and 4, exceeding MR-DCoT Zhang et al. (2026e) by 9.1% and 12.4% mAP, respectively. On night-sunny conditions, Table 5 shows that our GOLD reaches 60.5% mAP, a gain of 5.9% mAP, demonstrating strong generalization capability.

Cross-Camera. Table 6 reports the transfer from Cityscapes to BDD100K, where GOLD achieves 59.5% mAP and exceeds all the previous SOTA methods. Its strong performance on bus, car, motorcycle, and person suggests that using category prototypes as semantic anchors provides stable category-level guidance under changes in camera characteristics and scene appearance.

Adverse Weather. As shown in Table 7, GOLD achieves 65.6% mAP on Foggy Cityscapes, improving upon Bridge Hong et al. (2026) by 4.0% mAP and ranking first in seven of eight categories.

Table 8: Component ablation of GOLD on the four Diverse Weather target domains.

row	SCAdapter	SHEP	PAD	DF	DR	NR	NS	Avg.
1				42.3	49.2	28.7	51.2	42.9
2	✓			48.3	52.0	42.7	52.1	48.8
3	✓	✓		52.2	58.1	44.5	55.9	52.7
4	✓		✓	53.2	58.6	45.3	57.1	53.6
5	✓	✓	✓	**55.6**	**62.3**	**48.8**	**60.5**	**56.8**

Table 9: Ablation of the SCAdapter, backbone updating, and model scale.

row	Backbone	Update	Adaptation	DF	DR	NR	NS	Avg.
1	ConvNeXt-L	Frozen	Linear	52.3	58.9	44.8	58.1	53.5
2	ConvNeXt-L	Unfrozen	SCA	54.8	60.4	47.0	58.9	55.3
3	ConvNeXt-S	Frozen	SCA	54.3	60.7	46.2	58.6	54.9
4	ConvNeXt-B	Frozen	SCA	55.2	61.9	47.1	59.8	56.0
5	ConvNeXt-L	Frozen	SCA	**55.6**	**62.3**	**48.8**	**60.5**	**56.8**

Table 10: SHEP strength ablation. FG/CT/BG denote foreground/context/background; W/M/S denote weak/moderate/strong.

row	FG	CT	BG	DF	DR	NR	NS	Avg.
1	-	-	-	53.2	58.6	45.3	57.1	53.6
2	W	W	W	53.8	59.4	45.9	57.8	54.2
3	M	M	M	54.4	60.2	46.6	58.6	55.0
4	S	S	S	53.9	59.8	46.0	58.0	54.4
5	W	M	S	54.1	60.0	46.4	58.2	54.7
6	W	S	M	54.7	60.8	47.0	58.9	55.4
7	M	W	S	54.3	60.3	46.7	58.5	55.0
8	S	W	M	54.0	60.1	46.1	58.1	54.6
9	S	M	W	55.0	61.2	47.5	59.4	55.8
10	M	S	W	**55.6**	**62.3**	**48.8**	**60.5**	**56.8**

Table 11: PAD design ablation over prototype source, scales, and denoising stages. Aug. denotes the perturbed view.

row	Source	Scales	Stages	DF	DR	NR	NS	Avg.
1	-	-	-	52.2	58.1	44.5	55.9	52.7
2	Aug.	4	1–6	54.8	59.5	46.8	58.9	55.0
3	Clean	1	1-6	54.9	61.9	48.3	59.9	56.3
4	Clean	4	6	55.2	60.7	47.3	59.3	55.6
5	Clean	4	1–6	55.6	62.3	48.8	60.5	56.8



Figure 2: Visualization of heterogeneous spatial evidence decomposition.

Even under dense fog, our method can recover target domain objects from weakened foreground and context cues through heterogeneous perturbations and prototype-based semantic guidance.

5.4 Ablation Studies

Ablation of Components. Table 8 isolates the contributions of the three proposed modules. Compared with the baseline Chen et al. (2023) in *row 1*, introducing frozen DINOv3 features through the SCAdapter raises the average mAP from 42.9% to 48.8% in *row 2*, showing that spatial calibration makes the frozen foundation representations more suitable for iterative localization. Building on *row 2*, SHEP and PAD independently improve the average mAP by 3.9% and 4.8% mAP, respectively (*rows 3–4*). SHEP improves robustness by exposing proposal denoising to unreliable foreground, background, and context evidence, whereas PAD constrains proposal representations with clean category prototypes. Combining both modules achieves 56.8% mAP in *row 5*, indicating that perturbing localization evidence and preserving category semantics address complementary aspects of iterative proposal denoising.

Ablation of Spatial Calibration Adapter. Table 9 examines feature adaptation, backbone updating, and model scale. All other modules are enabled by default. With the same frozen ConvNeXt-L backbone, the Spatial Calibration Adapter improves the average mAP by 3.3% mAP over linear projection (*rows 1 and 5*). Keeping the backbone frozen yields a 1.5% mAP gain over end-to-end fine-tuning under the same adapter design (*rows 2 and 5*), indicating that preserving pretrained representations benefits cross-domain generalization. Scaling the frozen backbone from ConvNeXt-S to ConvNeXt-L further raises the average mAP from 54.9% to 56.8% (*rows 3–5*).

Ablation of Spatially Heterogeneous Evidence Perturbing. Table 10 compares uniform (*i.e., row 2-4*) and spatially heterogeneous perturbation-strength assignments(*i.e., row 5-10*). The best uniform setting applies moderate perturbation to all regions and achieves 55.0% mAP. By contrast, Denoising-Oriented Evidence Perturbation assigns moderate, strong, and weak perturbations to the foreground, localization context, and background, respectively, reaching 56.8% mAP (*row 10*). It surpasses no perturbation and the best uniform setting by 3.2% and 1.8% mAP, confirming the importance of perturbing each evidence region according to its role in proposal denoising.

Ablation of Prototype-Anchored Denoising. We examine prototype source, feature scales, and supervision stages of PAD in Table 11. The complete PAD improves the average mAP from 52.7% to 56.8% (compare *row 1* and *row 5*). Using clean (*row 5*) rather than perturbed views (*row 2*) for prototype construction yields a 1.8% mAP gain, suggesting that semantic anchors should be constructed from raw DINOv3 features of clean images. A comparison between *row 3* and *row 5* shows that

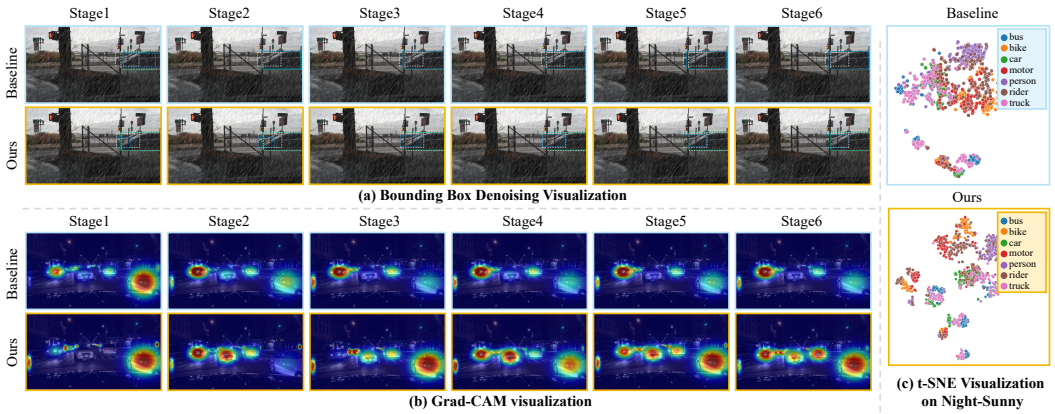

Figure 3: Qualitative comparison of the baseline model and GOLD in progressive box denoising, stage-wise attention, and feature distributions. More comparison in Appendix E, F.

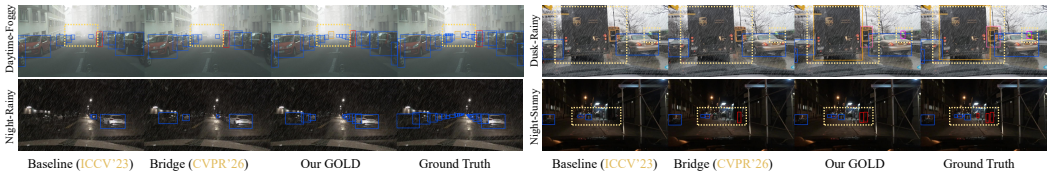

Figure 4: Qualitative detection results of the baseline model, Bridge, and GOLD, together with ground truth, across the four Diverse Weather target domains. More comparison in Appendix G.

multi-scale prototypes outperform single-scale prototypes by 0.5% mAP, demonstrating the value of complementary category information across feature levels. Applying prototype guidance across refinement stages improves upon final-stage-only supervision by 1.2% mAP (compare *row 4* and *row 5*), highlighting the benefit of preserving category semantics throughout iterative localization.

5.5 Qualitative Analysis

Stage-Wise Box Denoising. Figure 3(a) compares the six-stage denoising trajectories. Baseline model (DiffusionDet) misses the vehicle behind the leading car, whereas GOLD progressively recovers it from random proposals and produces an accurate bounding box.

Stage-Wise Attention Visualization. Figure 3(b) presents Grad-CAM maps across the six refinement stages. GOLD maintains stronger and more object-focused responses than baseline model, indicating more reliable visual evidence for proposal refinement.

Feature Distribution Visualization. Figure 3(c) shows t-SNE on Night-Sunny. Compared with DiffusionDet, GOLD yields more compact intra-class clusters and clearer inter-class separation, particularly for the commonly confused person and rider, and bicycle and motorcycle class pairs.

Detection Results. Figure 4 presents qualitative comparisons across the four unseen weather and illumination conditions in Diverse Weather. As highlighted by the yellow dashed boxes, GOLD detects challenging objects that are missed or incompletely localized by baseline and Bridge, producing more accurate and complete detections.

6 Conclusion

We presented **G**eneralized Pr**O**posa**L** **D**enoising (**GOLD**), a diffusion-based framework for domain-generalized object detection without access to target-domain data. GOLD integrates a Spatial Calibration Adapter for localization-sensitive feature calibration, Spatially Heterogeneous Evidence Perturbing for denoising-oriented domain diversification, and Prototype-Anchored Denoising for stage-wise semantic stabilization. Extensive experiments across diverse weather, camera, and adverse-weather shifts demonstrate consistent state-of-the-art performance and confirm the complementarity of the three components.

AI USE STATEMENT

We used GPT solely to assist with language polishing and improving the readability of the manuscript. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of this work.

REFERENCES

Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19830–19843, 2023.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Muhammad Sohail Danish, Muhammad Haris Khan, Muhammad Akhtar Munir, M Saquib Sarfraz, and Mohsen Ali. Improving single domain-generalized object detection: A focus on diversification and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17732–17742, 2024.

Xiaolong Guo, Chengxu Liu, Xueming Qian, Zhixiao Wang, Xubin Feng, and Yao Xue. Single-domain generalized object detection with frequency whitening and contrastive learning. *IEEE Transactions on Multimedia*, 27:6805–6818, 2025. doi: 10.1109/TMM.2025.3590915.

Boyong He, Yuxiang Ji, Zhuoyue Tan, and Liaoni Wu. Boosting domain generalized and adaptive detection with diffusion models: Fitness, generalization, and transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1912–1923, 2025a.

Boyong He, Yuxiang Ji, Qianwen Ye, Zhuoyue Tan, and Liaoni Wu. Generalized diffusion detector: Mining robust features from diffusion models for domain-generalized detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9921–9932, 2025b.

Mingbo Hong, Feng Liu, Caroline Gevaert, George Vosselman, and Hao Cheng. Bridge: Basis-driven causal inference marries VFMs for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31962–31973, 2026.

Han Jiang, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Diffusion-based source-biased model for single domain generalized object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1548–1557, 2025.

Junseok Lee, Sungho Shin, Seongju Lee, and Kyoobin Lee. Cd-fkd: Cross-domain feature knowledge distillation for robust single-domain generalization in object detection. *arXiv preprint arXiv:2603.16439*, 2026.

Wooju Lee, Dasol Hong, Hyungtae Lim, and Hyun Myung. Object-aware domain generalization for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2947–2955, 2024.

Chen Li, Huiying Xu, Changxin Gao, Zeyu Wang, Yun Liu, and Xinzhong Zhu. Towards single-source domain generalized object detection via causal visual prompts. In *Advances in Neural Information Processing Systems*, volume 38, pp. 104893–104921. Curran Associates, Inc., 2025.

Deng Li, Aming Wu, Yaowei Wang, and Yahong Han. Prompt-driven dynamic object-centric learning for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17606–17615, 2024.

Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8771–8780, 2021.

Yajing Liu, Shijun Zhou, Xiyao Liu, Chunhui Hao, Baojie Fan, and Jiandong Tian. Unbiased faster r-cnn for single-source domain generalized object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28838–28847, 2024.

Zhijie Rao, Jingcai Guo, Luyao Tang, Yue Huang, Xinghao Ding, and Song Guo. SRCD: Semantic reasoning with compound domains for single-domain generalized object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):12497–12506, 2024. doi: 10.1109/TNNLS.2024.3480120.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3219–3229, 2023.

Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 847–856, 2022.

Fan Wu, Jinling Gao, Lanqing Hong, Xinbing Wang, Chenghu Zhou, and Nanyang Ye. G-NAS: Generalizable neural architecture search for single domain generalization object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5958–5966, 2024.

Mingjun Xu, Lingyun Qin, Weijie Chen, Shiliang Pu, and Lei Zhang. Multi-view adversarial discriminator: Mine the non-causal factors for object detection in unseen domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8103–8112, 2023.

Xiaoran Xu, Jiangang Yang, Wenyue Chong, Wenhui Shi, Shichu Sun, Jing Xing, and Jian Liu. Boosting single-domain generalized object detection via vision-language knowledge interaction. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 131–140, 2025.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2633–2642. IEEE, 2020.

Peng Zhang, Xiang Yuan, and Gong Cheng. Unified interaction consistency learning for single-source domain-generalized object detection in urban scene. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12672–12680, 2026a.

Peng Zhang, Xiang Yuan, Cong Li, Junwei Han, and Gong Cheng. Once for all: An end-to-end paradigm for vlm-based domain-generalized object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6829–6838, 2026b.

Yupeng Zhang, Ruize Han, Ningnan Guo, Wei Feng, Song Wang, and Liang Wan. VFM^{4}SDG: Unveiling the power of VFMs for single-domain generalized object detection. *arXiv preprint arXiv:2604.21502*, 2026c.

Zihao Zhang, Aming Wu, and Yahong Han. Style evolving along chain-of-thought for unknown-domain object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14225–14234, 2025.

Zihao Zhang, Yang Li, Aming Wu, and Yahong Han. Simulating distribution dynamics: Liquid temporal feature evolution for single-domain generalized object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13061–13069, 2026d.

Zihao Zhang, Aming Wu, Yang Li, and Yahong Han. Prototype-anchored generalized manifold regression for unknown-domain object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026e. doi: 10.1109/TPAMI.2026.3714053. Early Access.

Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *Computer Vision – ECCV 2022*, pp. 535–552. Springer, 2022.

APPENDIX

A COMPARISON WITH DIFFUSION-BASED DETECTION METHODS

Diffusion has been used in object detection in two distinct ways. DiffusionDet models bounding boxes as noisy variables and progressively refines them into detections. GDD He et al. (2025b) and Boost He et al. (2025a) instead use pretrained image diffusion models to extract transferable visual features, which are subsequently processed by Faster R-CNN detection heads. GOLD follows DiffusionDet's box-denoising formulation, but addresses the reliability of the visual evidence and category semantics that guide each refinement stage. Table 12 summarizes these differences.

Diffusion for representations versus localization. GDD aggregates features extracted at multiple diffusion timesteps and transfers their knowledge to a conventional detector through feature- and object-level alignment. Boost improves this line of work with single-step feature extraction and an object-centered auxiliary branch conditioned on box masks and class prompts. Both methods exploit image-diffusion representations; their detection heads do not denoise bounding-box coordinates. GOLD instead initializes noisy box proposals and refines their positions and sizes over six detection stages. Thus, the diffusion process directly structures localization, while a frozen DINOv3 encoder supplies visual features.

Generalizing the denoising process. DiffusionDet already provides iterative box refinement, which GOLD retains. The additional challenge in DGOD is that the visual evidence used to update a proposal may change unevenly across foreground, object boundaries, and background. SHEP addresses this during source-only training by applying different perturbation strengths to foreground, localization context, and background. PAD constructs multi-scale category prototypes from clean, source-domain DINOv3 features and uses them to supervise matched proposal representations across the refinement stages. This differs from GDD's cross-detector knowledge transfer and Boost's consistency between ordinary and object-centered branches: the prototype constraint acts on the representations of proposals being denoised. SHEP and prototype supervision are training-time components; inference retains the calibrated backbone, feature pyramid, and DiffusionDet head.

Empirical comparison. Table 13 reproduces the Diverse Weather results reported in the main paper. GOLD obtains 56.8% mAP on average, exceeding the reported DiffusionDet, GDD, and Boost entries by 13.9%, 16.6%, and 15.0% mAP, respectively. The component ablation provides

Table 12: How diffusion is used in detection. DiffusionDet Chen et al. (2023) denoises boxes, whereas GDD He et al. (2025b) and Boost He et al. (2025a) extract features from image diffusion models.

Method	Diffusion applied to	Box localization	Generalization strategy
DiffusionDet	Box coordinates	Iterative box denoising	Generic detection; no DG-specific guidance
GDD	Image features across multiple steps	RPN and RoI regression	Diffusion-feature extraction and feature/object transfer
Boost	Image features at a single step	RPN and RoI regression	Object-centered auxiliary branch and branch consistency
GOLD	Box coordinates	Iterative box denoising	Spatially calibrated features, region-wise perturbation, and stage-wise prototype guidance

Table 13: Diverse Weather DG results (% mAP) reproduced from the main paper. DF, DR, NR, and NS denote daytime-foggy, dusk-rainy, night-rainy, and night-sunny. Avg. is their arithmetic mean.

Method	DF	DR	NR	NS	Avg.
DiffusionDet	42.3	49.2	28.7	51.2	42.9
GDD	43.3	42.5	27.8	47.0	40.2
Boost	45.9	44.3	28.2	48.8	41.8
GOLD	**55.6**	**62.3**	**48.8**	**60.5**	**56.8**

Table 14: Hyperparameter settings used in GOLD.

Symbol	Description	Value
Training and architecture		
$\{d_l\}_{l=2}^{5}$	Bottleneck dimensions of SCA at the four backbone levels	$(64, 128, 256, 512)$
N_{iter}	Number of training iterations	450 000
B	Training batch size	4
lr_0	Initial learning rate	2.5×10^{-5}
Spatially Heterogeneous Evidence Perturbing		
(s^-, s^+)	Sampling range for the foreground-support scale	$(0.5, 1.5)$
η	Maximum normalized center displacement	0.4
(λ^-, λ^+)	Sampling range for the context-expansion ratio	$(0.2, 0.7)$
$d_{\min}$	Minimum size of the foreground-support region	24 pixels
$m_{\min}$	Minimum context-expansion margin	12 pixels
ρ_{fg}	Residual strength for the foreground region	$\mathcal{U}(0.58, 0.78)$
ρ_{ctx}	Residual strength for the localization-context region	$\mathcal{U}(0.84, 0.96)$
ρ_{bg}	Residual strength for the background region	$\mathcal{U}(0.30, 0.52)$
$\mathcal{T}$	Perturbation operations	color jittering, Gaussian blur, and Gaussian noise
Prototype-Anchored Denoising		
μ	EMA momentum used to update the prototype bank	0.95
τ	Temperature of the prototype contrastive objective	0.07
λ_{proto}	Weight of the prototype-guided denoising loss	0.2
$\boldsymbol{\omega}$	Loss weights for the six proposal-refinement stages	$(0.05, 0.05, 0.10, 0.20, 0.25, 0.40)$
$\boldsymbol{\beta}$	Loss weights for the four prototype feature levels	$(0.25, 0.25, 0.25, 0.25)$

a more direct view of the changes to the box-denoising baseline: the average rises from 42.9% to 48.8% mAP with SCAdapter, and reaches 56.8% mAP when SHEP and PAD are both enabled.

B Hyperparameter Settings

Table 14 lists the settings used for the detector, spatial evidence perturbation, and prototype-guided training. The three region-specific strength ranges implement the ordering $\rho_{\text{bg}} < \rho_{\text{fg}} < \rho_{\text{ctx}}$ in SHEP. The stage and feature-level weights in PAD control how the clean semantic references contribute to proposal supervision.

Algorithm 1 One training iteration of GOLD.

Require: Source mini-batch $\mathcal{B}$; frozen DINOv3 encoder F; trainable parameters Θ; prototype bank $\mathcal{P}$
Ensure: Updated Θ and $\mathcal{P}$
Spatial evidence perturbation
Construct foreground, localization-context, and background regions from source ground-truth boxes
Apply region-dependent perturbations to obtain $\widetilde{\mathcal{B}} = \mathrm{SHEP}(\mathcal{B})$
Clean prototype construction
Extract multi-scale ground-truth RoI features from the clean images using F
Update the class-wise prototype bank $\mathcal{P}$ with stop-gradient EMA
Prototype-anchored proposal denoising
Encode $\widetilde{\mathcal{B}}$ with F and the SCAdapter; initialize noisy box proposals
for $s = 1, \ldots, 6$ **do**
    Refine proposals and compute the stage detection loss
    Match proposals to ground truth and apply prototype supervision to positive matches
end for
Update Θ using $\mathcal{L}_{\mathrm{GOLD}} = \mathcal{L}_{\mathrm{det}} + \lambda_{\mathrm{proto}}\mathcal{L}_{\mathrm{proto}}$
Inference: Use F, the SCAdapter, FPN, and DiffusionDet head.

Table 15: Training and inference efficiency with matched DINOv3 backbones. Training throughput is measured in iterations per second (it/s); inference throughput is measured in images per second (img/s). GPU memory is reported in GB.

	Training		Inference	
Method	Throughput ↑	GPU memory ↓	Throughput ↑	GPU memory ↓
DINOv3-T (Baseline)	2.13	10.13	14.27	3.91
DINOv3-T (GOLD)	1.49	12.39	14.27	3.91
DINOv3-S (Baseline)	1.91	11.32	11.58	4.37
DINOv3-S (GOLD)	1.34	14.28	11.58	4.37
DINOv3-B (Baseline)	1.73	12.71	8.71	5.36
DINOv3-B (GOLD)	1.12	15.53	8.71	5.36
DINOv3-L (Baseline)	1.54	14.18	6.60	6.42
DINOv3-L (GOLD)	0.95	17.62	6.60	6.42

C Training Procedure

Algorithm 1 summarizes one source-domain training iteration. SHEP first creates a perturbed view while preserving the original box annotations. Clean ground-truth RoI features update the prototype bank without gradients; the perturbed view then drives the six-stage detection loss and the stage-weighted prototype objective. At inference, neither the perturbation operation nor prototype supervision is required.

D Training Efficiency Analysis

We compare GOLD with the baseline under matched DINOv3 backbone sizes in the Cityscapes→BDD100K setting. Table 15 reports training throughput and GPU memory, alongside inference measurements for the same model pairs.

GOLD incurs additional training cost across all four backbone sizes: relative to the matched baselines, throughput decreases by 29.8–38.3%, while GPU memory increases by 2.26–3.44 GB. For the DINOv3-L configuration used in our main experiments, training throughput changes from 1.54 to 0.95 it/s, and memory usage changes from 14.18 to 17.62 GB. SHEP and the clean-view prototype construction and supervision in PAD add operations during training, consistent with this increase in cost.

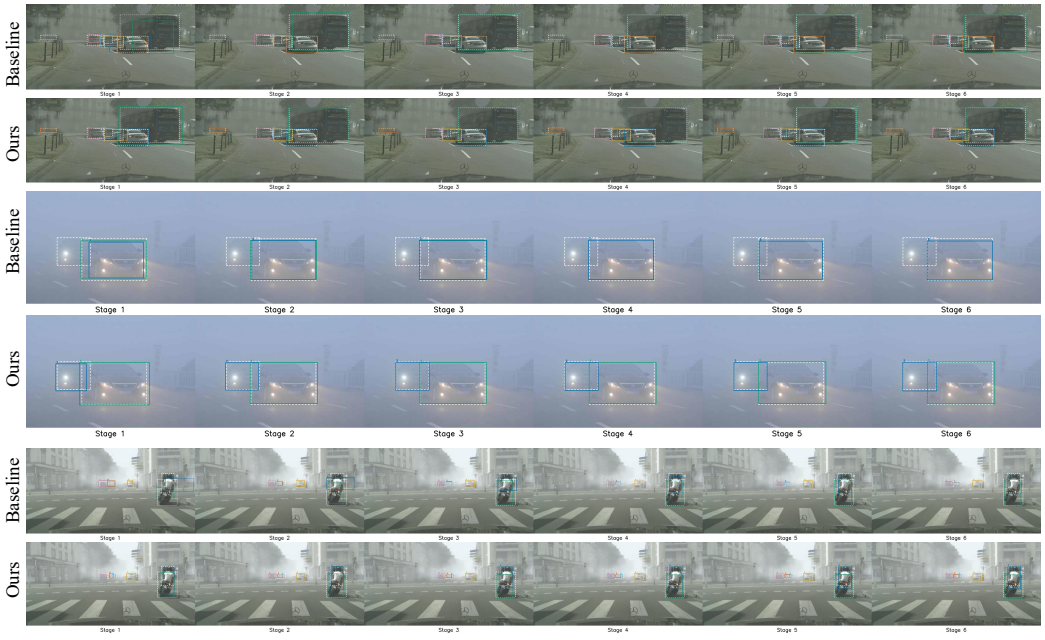

Figure 5: More bounding box denoising visualizations on Daytime-Foggy.

At the reported precision, each GOLD model has the same inference throughput and GPU memory as its matched baseline. SHEP and the PAD-specific supervision branch are removed at inference, so their training cost does not carry over to the measured inference resource usage in this setting.

E Stage-Wise Bounding-Box Denoising

Figures 5–8 provide additional six-stage proposal-refinement examples on Daytime-Foggy, Dusk-Rainy, Night-Rainy, and Night-Sunny. Each example places the DiffusionDet baseline and GOLD on adjacent rows for the same target-domain image. The baseline performs box denoising without the spatial calibration, region-dependent perturbation, or prototype supervision used to train GOLD. The columns show how their predictions evolve from the first to the sixth detection stage.

Under fog and rain, distant objects and object boundaries provide weak localization cues, while nighttime headlights and reflections can dominate the visible evidence. In these examples, the baseline sometimes retains incomplete or poorly placed boxes through the later stages and misses dim instances. GOLD more often preserves candidate objects and progressively adjusts the boxes toward their visible extents. The Night-Rainy examples in Fig. 7 make the difficulty especially apparent: much of the scene is dark, and the available cues are concentrated around vehicle lights. The comparison illustrates the benefit of training the refinement process with heterogeneous spatial evidence while maintaining category-level guidance from clean prototypes.

F Stage-Wise Attention Visualization

Figure 9 extends the Grad-CAM comparison to additional examples across the four Diverse Weather target domains. Each baseline–GOLD pair is shown over the six refinement stages. Relative to the baseline, GOLD generally maintains more distinct responses around candidate objects, including visually weak objects under poor illumination and weather. The baseline responses can be concentrated on the most salient bright regions or become diffuse as refinement proceeds. These maps provide a qualitative view of the visual evidence used by the two detectors.

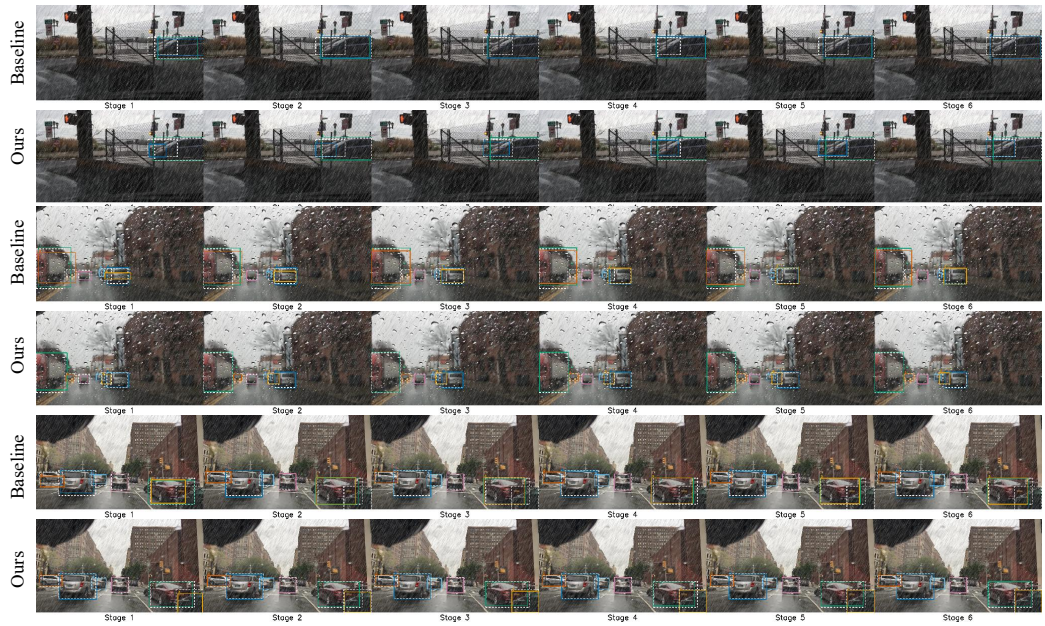

Figure 6: More bounding box denoising visualizations on Dusk-Rainy.

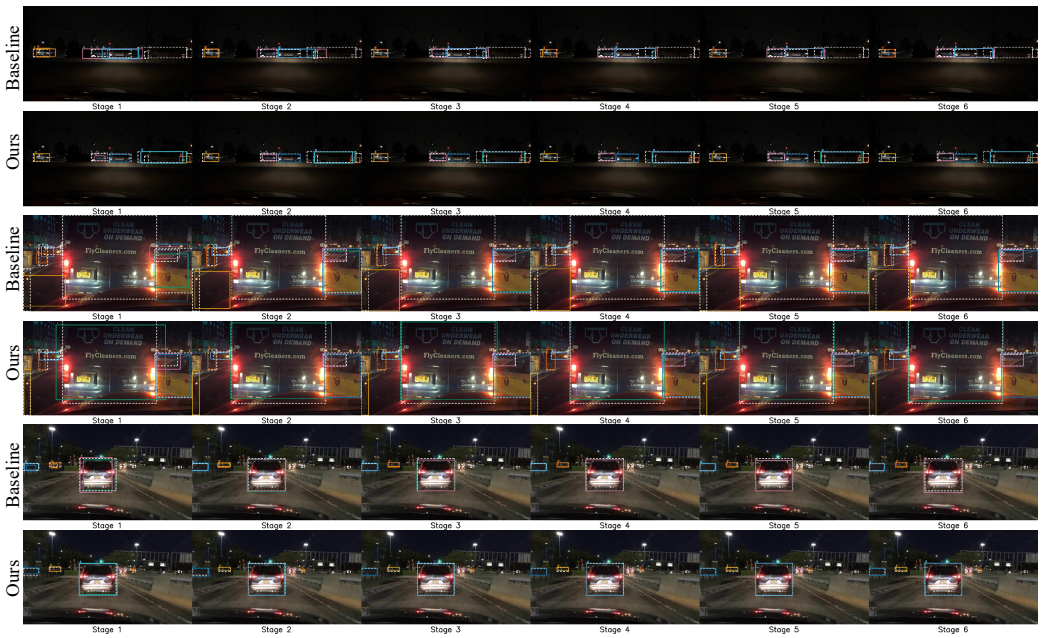

Figure 7: More bounding box denoising visualizations on Night-Rainy.

G Qualitative Comparison with Other Detectors

Figures 10–13 compare the predictions of DiffusionDet, Bridge, and GOLD with the ground-truth annotations on the four unseen weather domains. Compared with the source-only DiffusionDet baseline, GOLD is trained with a spatially calibrated frozen DINOv3 encoder, SHEP, and PAD, while retaining iterative box refinement at test time. Bridge provides a second, strong DGOD comparison

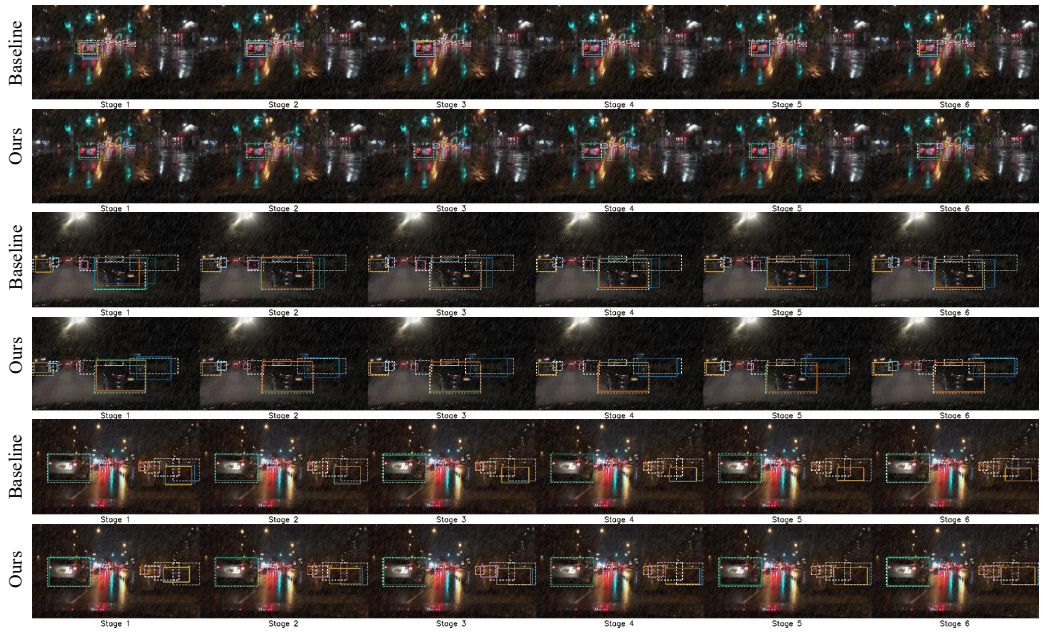

Figure 8: More bounding box denoising visualizations on Night-Sunny.

with a different approach to generalization. The columns in each figure use the same image so that missed objects and localization differences can be inspected directly.

In Daytime-Foggy, weak contrast makes distant vehicles and road users difficult to separate from the background. For example, the sixth scene of Fig. 10 shows additional road users recovered by GOLD that the DiffusionDet baseline misses. The Dusk-Rainy examples in Fig. 11 include wet roads and partially obscured vehicles; GOLD retains detections for prominent objects while recovering more of the smaller objects in the scene. The difference is most visible under Night-Rainy conditions. In the fourth scene of Fig. 12, several dim vehicles along the road are absent from the baseline output, whereas GOLD produces more detections consistent with the annotations. Night-Sunny scenes in Fig. 13 further show that this behavior extends beyond rain; for instance, GOLD recovers the roadside vehicle in the fourth scene that is missed by the baseline. Across these examples, GOLD produces more complete detections under degraded visual evidence, in line with the quantitative results reported in the main text.

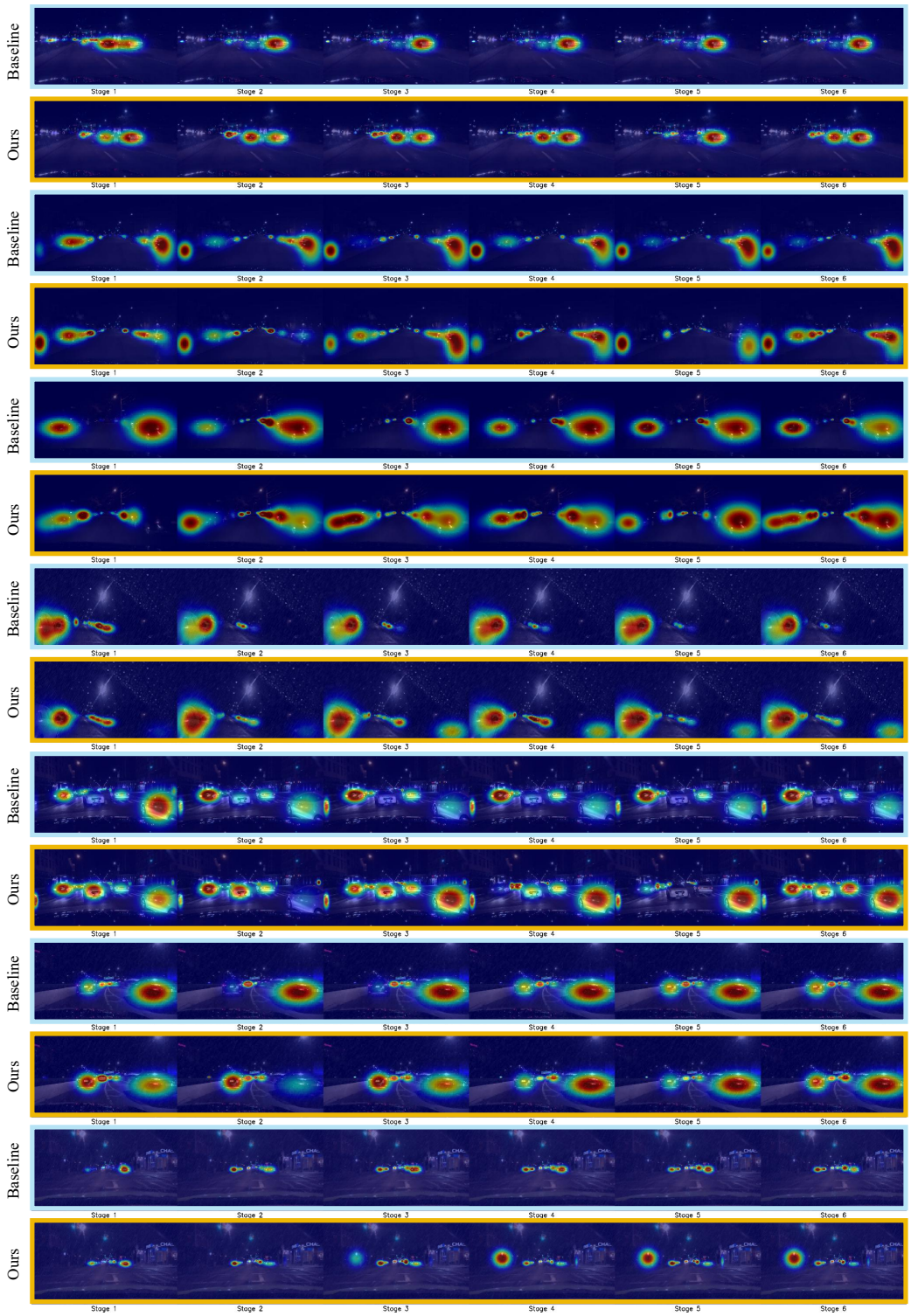

Figure 9: More Grad-CAM visualization of our method and the baseline on Diverse Weather.

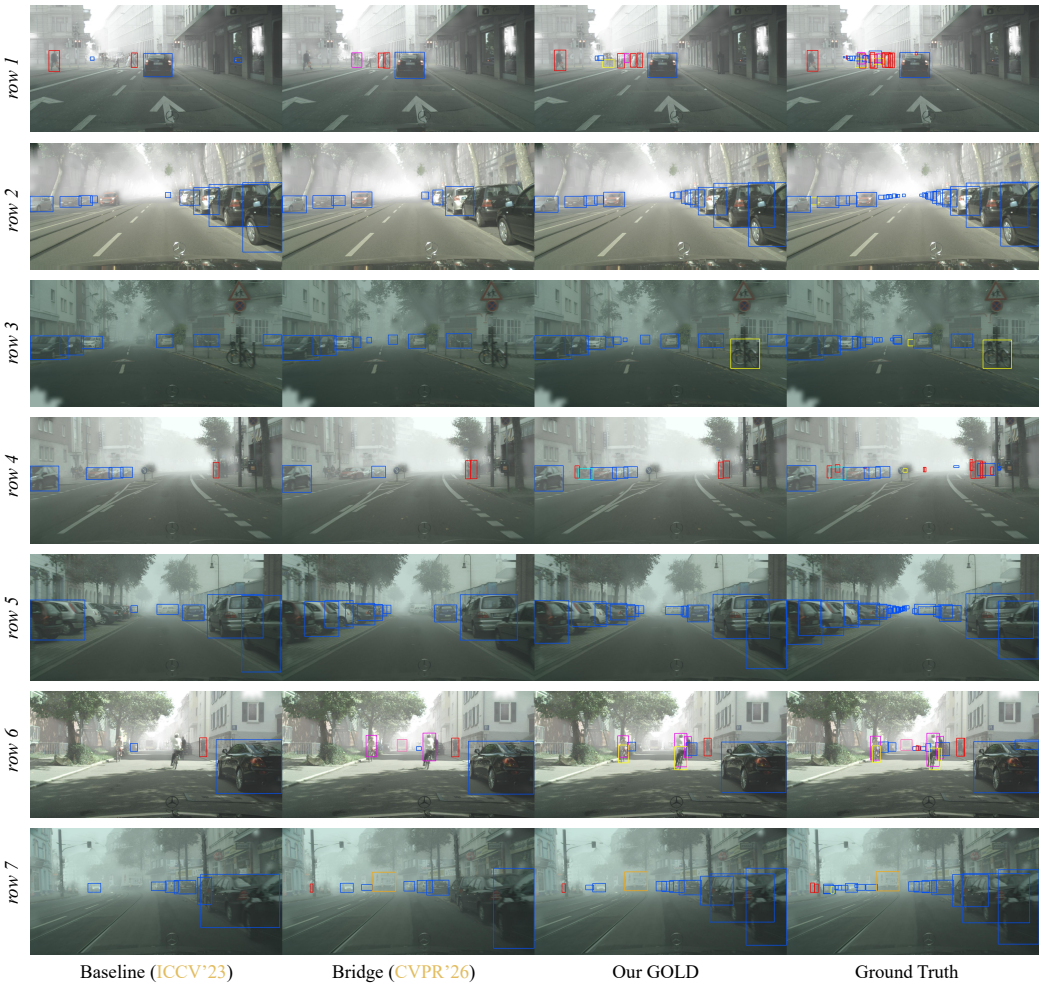

Figure 10: More qualitative detection results on Daytime-Foggy.



Figure 11: More qualitative detection results on Dusk-Rainy.



Figure 12: More qualitative detection results on Night-Rainy.



Figure 13: More qualitative detection results on Night-Sunny.