MIXTURE OF PROTOTYPE EXPERTS WITH PREFERENCE-GUIDED ROUTING FOR SINGLE DOMAIN GENERALIZED OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single Domain Generalized Object Detection (SDGOD) trains a detector on a single labeled source domain for generalization to unseen domains. Existing approaches learn invariant features or broaden the source distribution, but limited source diversity prevents them from modeling complex combinations of unseen visual conditions. We propose Mixture of Prototype Experts (MoPE), a framework that factorizes visual conditions into atomic prototypes across photometric, textural, and spatial dimensions, each paired with a corresponding expert. To learn input-adaptive expert composition, we formulate routing as a preference learning problem over multiple coexisting visual attributes, enabling the router to assign higher selection probabilities to scene-relevant prototype experts than to nonrelevant ones. The selected expert outputs are then sparsely aggregated with a shared expert for multi-scale feature fusion. This preference learning is grounded in a Multi-Style Augmentation (MSA) strategy that broadens illumination, weather, and appearance variations without target-domain information while preserving object content and geometry, ensuring the router observes sufficient diversity for learning generalizable routing preferences. Extensive experiments demonstrate consistent generalization gains across unseen weather conditions, artistic styles, and different backbone architectures while maintaining a favorable accuracy-efficiency trade-off. *Code and data will be publicly available.*

1 INTRODUCTION

Object detectors perform strongly in applications such as autonomous driving and video surveillance, but their performance often degrades under distribution shifts caused by variations in illumination, weather, and background. Domain Generalization (DG) (Wang et al., 2021; Jin et al., 2022) aims to learn transferable features from multiple source domains without access to target-domain data, commonly through data augmentation and feature disentanglement. DG has been studied in autonomous driving (Li et al., 2023; Choi et al., 2021), crowd counting (Chen et al., 2025; Hu et al., 2026b), and open-vocabulary detection (Jiao et al., 2025; Hu et al., 2026a). Single Domain Generalized Object Detection (SDGOD) (Wu & Deng, 2022) adopts a stricter setting in which a detector is trained on a single labeled source domain and evaluated on unseen domains for both classification and localization. Without cross-domain diversity during training, the detector must learn transferable features from a single distribution, making SDGOD particularly challenging and practically important.

Existing SDGOD approaches broadly follow two complementary directions: **invariant feature learning**; **semantic and distribution modeling** (Fig. 1(a)). Invariant feature learning methods (Lee et al., 2024; Guo et al., 2025) seek features that remain stable across domains by suppressing source-specific factors. However, limited source-domain support can leave residual bias toward the source distribution (Xu et al., 2023; Liu et al., 2024). Semantic and distribution modeling methods (Zhang et al., 2025; 2026; Jiang et al., 2025) incorporate semantic priors or broaden the observed visual variations through prompt guidance, causal intervention, style transformation, feature recomposition, and generative augmentation. Although these strategies improve robustness, the diverse combinations of visual factors encountered in unseen domains remain difficult to cover from a single source distribution.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

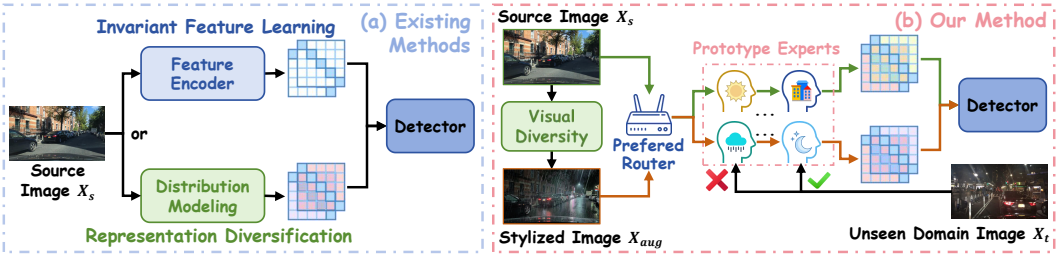

Figure 1: Comparison of SDGOD paradigms. (a) Existing methods rely on invariant feature learning or source distribution diversification. (b) Our method combines visual diversification with preference-guided expert routing for input-adaptive feature composition.

Complementing these directions, we observe that Mixture of Experts (MoE) (Jacobs et al., 1991; Fedus et al., 2022) provides a natural paradigm for generalization across heterogeneous distributions by combining specialized experts through dynamic routing. However, directly applying MoE is insufficient: when the router is optimized solely by the detection objective, it may repeatedly favor a small subset of experts across diverse inputs, causing routing collapse and reducing the expert ensemble to a nearly static computation path. We therefore propose MoPE, a Mixture of Prototype Experts framework tailored to SDGOD. MoPE characterizes visual conditions using predefined atomic prototypes along the photometric, textural, and spatial dimensions, with each prototype associated with a corresponding expert. A router trained through multi-label preference optimization dynamically selects and aggregates multiple experts for each input, as illustrated in Fig. 1(b). Analogous to how weighted RGB basis colors span a rich color space, the weighted combination of prototype experts produces diverse, input-adaptive feature representations for coexisting scene attributes.

Specifically, MoPE builds on the YOLO detector (Jocher & Qiu, 2024) by inserting prototype expert modules into the P3, P4, and P5 backbone stages while retaining the detector neck and head. The framework comprises three components. First, we optimize the router with Multi-Label Preference Optimization (MPO), which associates each input with multiple compatible semantic prototypes. Second, sparse routing activates and weights the corresponding prototype experts, whose aggregated outputs are combined with a shared expert. Third, MSA broadens the source distribution by generating variations in illumination, weather, and appearance while preserving object content and geometry, providing diverse observations for routing and detection.

Our contributions are threefold. (1) We propose MoPE, a prototype-expert framework for SDGOD that factorizes visual conditions into predefined atomic prototypes along the photometric, textural, and spatial dimensions. MoPE dynamically composes the corresponding experts to produce input-adaptive feature representations. (2) We formulate expert routing as a multi-label preference optimization problem, in which the router ranks multiple scene-relevant prototype experts above rejected experts and performs sparse activation with shared-expert aggregation. (3) Extensive experiments across weather and artistic benchmarks and multiple backbones demonstrate consistent generalization gains with a favorable accuracy-efficiency trade-off.

2 Related Work

Single domain generalized object detection. Early methods primarily perform **invariant feature learning.** For example, DIDN (Lin et al., 2021) disentangles domain-invariant and domain-specific features, while S-DGOD (Wu & Deng, 2022) combines cyclic disentanglement with self-distillation. FWCL (Guo et al., 2025) further employs frequency-domain whitening and contrastive learning to suppress domain-specific information. Nevertheless, limited source-domain support can bias learned features toward the source distribution (Xu et al., 2023). Recent studies improve generalization through **Semantic and distribution modeling.** For instances, CLIPGap (Vidit et al., 2023) incorporates vision-language prompts, UFR (Liu et al., 2024) models domain shifts through causal intervention, and Prompt-D (Li et al., 2024) uses scene-level descriptions to guide gated predictions. SECoT (Zhang et al., 2025) progressively integrates and expands style information through a chain-of-thought style modeling process to capture complex visual variations. F-KAN (Zhang et al., 2026) decomposes image features into Fourier amplitude and phase components and recombines them to

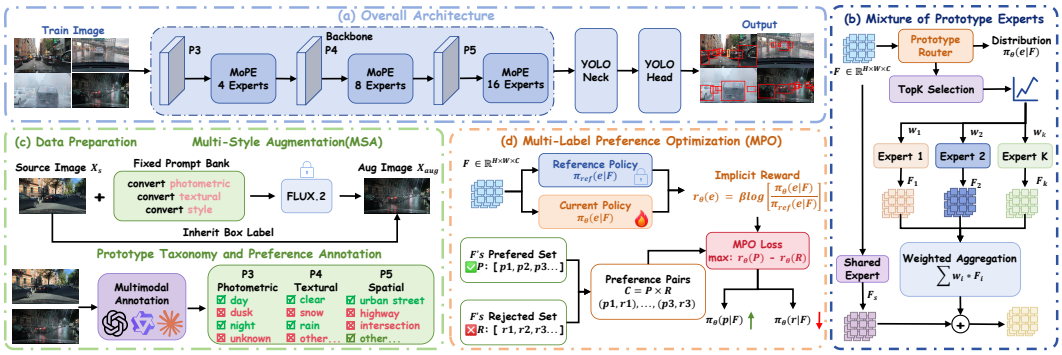

Figure 2: Overview of MoPE. (a) MoPE modules are integrated into P3–P5. (b) The router sparsely selects prototype experts and combines them with a shared expert. (c) MSA enriches source-domain visual conditions, while multimodal annotation constructs image-specific prototype preferences. (d) MPO increases the relative probabilities of preferred prototype experts.

generate samples with both style and content diversity. SDG-DiffDet (Jiang et al., 2025) models source-to-augmented distribution shifts using diffusion.

Mixture of Experts (MoE) models combine multiple expert networks through an input dependent router, allowing different experts to specialize in distinct regions of the input distribution (Jacobs et al., 1991). Sparse MoE (Shazeer et al., 2017), GShard (Lepikhin et al., 2020), and Switch Transformer (Fedus et al., 2022) activate only a subset of experts, increasing model capacity without a proportional increase in computation. ST-MoE (Zoph et al., 2022) further improves training stability and transferability. YOLO-ESMoE (Lin et al., 2026) extends sparse expert routing to object detection. However, task-driven routing alone does not explicitly associate experts with visual conditions, leaving specialization underconstrained and making routing collapse a potential failure mode. Building on MoE's ability to model heterogeneous distributions, we introduce it into SDGOD and formulate expert routing as a multi-label prototype preference optimization problem.

3 Proposed Method

3.1 Problem Statement

Let $\mathcal{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$ denote a labeled source domain dataset, where x_s^i is the i-th image and $y_s^i = \{c_i, b_i\}$ contains its category label c_i and bounding box coordinates b_i. We further consider T unseen target domains $\{\mathcal{D}_t^j\}_{j=1}^{T}$, where each $\mathcal{D}_t^j = \{x_{t,j}^i\}_{i=1}^{N_t^j}$ consists of unlabeled images. All domains share the same category label space. In the Single Domain Generalized Object Detection (SDGOD) setting, the model learns solely from the labeled source domain $\mathcal{D}_s$ without access to any target-domain data. The goal is to learn a robust detector that generalizes to multiple unseen target domains $\{\mathcal{D}_t^j\}_{j=1}^{T}$.

Overview. The proposed MoPE factorizes visual conditions into atomic prototypes, each mapped to a dedicated expert, and supervises routing with image-level prototype preferences for input-adaptive expert composition. As illustrated in Fig. 2, MoPE forms a unified pipeline comprising three closely coupled stages. **Data Preparation (Sec. 3.2)** defines the multi-granularity prototype taxonomy and employs Multi-Style Augmentation (MSA) to broaden the source-domain visual distribution. **Preference Learning for Prototype Routing (Sec. 3.3)** constructs image-specific prototype preferences and uses Multi-Label Preference Optimization (MPO) to optimize independently parameterized routers at P3, P4, and P5. Given an input feature, each router predicts a distribution over its stage-specific experts, where every output coordinate corresponds to a fixed prototype, i.e., expert index. **Routing Execution and Training (Sec. 3.4)** converts these distributions into sparse Top-K_l expert activation, aggregates the selected outputs with a shared expert, and passes the resulting multi-scale features to the detector neck and head for localization and classification. Together, these stages connect semantic data preparation, preference-supervised routing, and sparse expert execution, enabling input-adaptive expert composition without accessing target-domain data.

3.2 Data Preparation

MoPE requires both a target-independent label space for organizing expert specialization and sufficiently diverse visual observations for learning meaningful routing preferences. We therefore construct a fixed, multi-granularity taxonomy of photometric, textural, and spatial prototypes, with each prototype index assigned to one expert. The taxonomy supplies semantic labels for preference annotation, while offline Multi-Style Augmentation uses its photometric and textural dimensions to broaden the illumination and weather conditions observed during training. Together, the taxonomy and augmented data provide the semantic supervision and visual diversity required for subsequent preference learning.

3.2.1 Prototype Taxonomy

At stage $l \in \{3, 4, 5\}$, we define the semantic prototype space as

$$\mathcal{Q}^l = \{q_e^l\}_{e=1}^{E_l}, \qquad q_e^l \longleftrightarrow \mathrm{Expert}_e^l, \tag{1}$$

where each q_e^l is a discrete semantic category (e.g., "daytime", "rainy", "highway") with a textual label, and its index e identifies the corresponding expert through a fixed one-to-one mapping. The prototypes are categorical rather than metric: they serve as semantic anchors in the label space for preference annotation, while the router and expert networks operate on visual features in the feature space. During training, the taxonomy and the prototype–expert mapping remain fixed; only the router parameters and expert weights are updated. This fixed mapping ensures that each expert is consistently associated with a specific semantic attribute throughout training, providing a stable target for preference supervision.

We organize $(E_3, E_4, E_5) = (4, 8, 16)$ prototypes in a coarse-to-fine hierarchy aligned with the progressive depth of the backbone. This design follows the established feature hierarchy in CNNs, where shallow layers are most sensitive to low-level photometric attributes (illumination and color), intermediate layers capture textural patterns (surface and atmospheric conditions), and deep layers encode high-level semantic structures (scene layout and spatial context) (Zeiler & Fergus, 2014; Raghu et al., 2021). Accordingly, P3 describes coarse photometric factors, P4 describes intermediate textural attributes, and P5 describes fine-grained spatial layouts. Specifically, P3 contains four mutually exclusive photometric prototypes: *daytime*, *nighttime*, *dusk*, and *unknown*. P4 contains eight textural prototypes covering weather and surface appearance conditions (e.g., *clear*, *rainy*, *snowy*, *foggy*, etc.), where multiple prototypes may coexist in an image. P5 contains sixteen spatial prototypes, including *urban street*, *highway*, *intersection*, etc., and likewise permits multiple coexisting attributes. The complete taxonomy and expert-index assignments are provided in Appendix F.

3.2.2 Multi-Style Augmentation (MSA)

We now leverage the prototype taxonomy established in Sec. 3.2.1 to diversify the training distribution via style generation. For prompt construction, we use only the photometric (P3) and textural (P4) prototypes, as these directly govern illumination and surface appearance, the attributes most amenable to image-level style editing. The spatial prototypes (P5) are intentionally excluded from augmentation prompts to preserve the original scene geometry and layout, ensuring the generated images remain structurally aligned with their source annotations. We build a fixed prompt bank by combining P3 photometric prototypes (e.g., daytime, nighttime) with P4 textural prototypes (e.g., clear, rainy, snowy), covering a broad spectrum of temporal and weather conditions without targeting any specific unseen domain. Before offline generation, we select a set of prompts from this bank and randomly assign the selected prompts to the source images. Each image is then translated once using the pretrained diffusion model FLUX.2 (without fine-tuning).

For a labeled source image $(x_s, y_s^{\mathrm{det}})$, the generated image is $x_{\mathrm{aug}} = \mathrm{FLUX}(x_s, \textit{prompt})$. Since the generation preserves object categories, scales, viewpoint, and spatial arrangement, each augmented image inherits the original bounding box annotations $y_{\mathrm{aug}}^{\mathrm{det}} = y_s^{\mathrm{det}}$. The prompt bank and assignment are fixed before training and use no target-domain data or statistics. Both source and generated images are subsequently fed into the preference annotation procedure (Sec. 3.3.1), providing the routing policy with diverse photometric and textural observations while keeping spatial structure intact. Appendix G provides the complete prompt bank and analyzes the consistency and visual quality of the generated images.

3.3 Preference Learning for Prototype Routing

With the prototype taxonomy and augmented data in place, the next question is how to teach the router to select appropriate experts for each input. The fixed prototype–expert mapping established in Sec. 3.2.1 provides the key: each prototype is an index that identifies a specific expert, so selecting compatible prototypes for an image is equivalent to specifying which experts should be favored. The core idea, then, is to convert these semantic prototype choices into explicit supervisory signals for routing decisions. This section proceeds in two steps. First, we assign each training image a set of preferred and rejected prototype indices via multimodal LLM annotation, establishing what the router should favor or avoid. Second, we optimize the policy router with Multi-Label Preference Optimization (MPO) against a frozen reference policy, extending standard DPO to handle multiple coexisting preferences and explicitly ranking all preferred experts above rejected ones. Together, these two steps transform the static prototype taxonomy into a dynamic, learnable routing preference that generalizes across diverse visual conditions.

3.3.1 Multi-Granularity Preference Construction

Preference annotation. Given the fixed taxonomy, we employ a multimodal large language model to analyze each source or generated training image and identify the prototypes that best describe its visual content across P3, P4, and P5, thereby providing multi-granularity preference supervision over photometric, textural, and spatial attributes. Let $\mathcal{E}^l = \{1, \dots, E_l\}$ denote the expert-index set at stage $l \in \{3, 4, 5\}$. For the b-th image x_b, the preferred set is $\mathcal{P}_b^l = \{e \in \mathcal{E}^l \mid q_e^l \text{ describes an attribute present in } x_b\}$. The annotation output is the structured tuple $(\mathcal{P}_b^3, \mathcal{P}_b^4, \mathcal{P}_b^5)$. Because q_e^l and Expert_e^l share the same index, $\mathcal{P}_b^l$ identifies both the selected semantic categories and the experts preferred for routing; the remaining indices form $\mathcal{R}_b^l = \mathcal{E}^l \setminus \mathcal{P}_b^l$. This automated process uses only the training images and predefined taxonomy, requiring neither manual preference annotation nor target-domain data or statistics.

Design logic. The preference format follows the semantic structure of the taxonomy. At P3, photometric conditions are mutually exclusive; an image cannot be simultaneously daytime and nighttime, hence $|\mathcal{P}_b^3| = 1$. At P4 and P5, however, an image may exhibit multiple coexisting attributes, so we allow multiple preferred prototypes. Standard DPO assumes a single preferred–rejected pair and cannot directly handle such multi-label cases. We therefore define the comparison set $\mathcal{C}_b^l = \mathcal{P}_b^l \times \mathcal{R}_b^l$, pairing every preferred expert with every rejected expert.

3.3.2 Multi-Label Preference Optimization (MPO)

We adopt a two-stage training strategy. In the first T_c epochs, we train the detector with the standard detection loss only, allowing the router to learn a reasonable initial routing behavior from detection supervision. After this warm-up, we freeze the backbone and its routers as the reference policy, and activate MPO to explicitly shape routing preferences. This ensures that MPO starts from a stable, task-aware initialization rather than from random noise.

Router architecture. At each stage $l \in \{3, 4, 5\}$, the policy router takes the stage feature $\mathbf{F}_b^l$ as input and comprises a lightweight convolutional encoder g_θ^l followed by an expert-score projection Proj^l. The encoder produces an image-level routing representation $\mathbf{H}_b^l = g_\theta^l(\mathbf{F}_b^l)$, which the projection maps to E_l routing logits $\mathbf{z}_b^l = \mathrm{Proj}^l(\mathbf{H}_b^l) \in \mathbb{R}^{E_l}$. Applying Softmax across the E_l experts gives the policy distribution: $\pi_\theta^l(e|\mathbf{F}_b^l) = \exp(z_{b,e}^l) / \sum_{j=1}^{E_l} \exp(z_{b,j}^l)$. We define the reference policy f_{ref} as the frozen copy of the complete backbone, including all prototype experts and policy routers, at epoch T_c. For a given input x_b, $f_{\mathrm{ref}}^l(x_b)$ produces the stage feature $\mathbf{F}_{\mathrm{ref},b}^l$ and the reference routing distribution π_{ref}^l via the same architecture described above. This fixed reference remains unchanged during MPO training, serving as the anchor for computing implicit rewards. After f_{ref} is constructed, the original detector f_θ remains trainable as the current policy, and its stage-l router produces the routing distribution π_θ^l. The router is jointly optimized by the detection objective and MPO, with the latter increasing the relative routing probabilities of preferred prototype experts.

Multi-label preference optimization (MPO). The reference policy f_{ref} introduced above independently produces the stage feature $\mathbf{F}_{\mathrm{ref},b}^l = f_{\mathrm{ref}}^l(x_b)$ and the corresponding routing distribution π_{ref}^l. We define the implicit reward of expert e from the current and reference routing probabili-

ties: $\hat{r}_\theta^{b,l}(e) = \beta \log(\pi_\theta^l(e|\mathbf{F}_b^l)/\pi_{\text{ref}}^l(e|\mathbf{F}_{\text{ref},b}^l))$. The implicit reward measures the relative change in the routing probability of expert e. It is positive when the current policy assigns a higher probability than the reference policy, negative when the probability decreases, and zero when the two probabilities are equal. For each pair $(p, r) \in \mathcal{C}_b^l$, the current-policy preference margin is $\Delta_\theta^{b,l}(p, r) = \log \pi_\theta^l(p|\mathbf{F}_b^l) - \log \pi_\theta^l(r|\mathbf{F}_b^l)$. And the corresponding reference-policy margin is $\Delta_{\text{ref}}^{b,l}(p, r) = \log \pi_{\text{ref}}^l(p|\mathbf{F}_{\text{ref},b}^l) - \log \pi_{\text{ref}}^l(r|\mathbf{F}_{\text{ref},b}^l)$. Their difference gives the implicit reward margin:

$$\hat{r}_\theta^{b,l}(p) - \hat{r}_\theta^{b,l}(r) = \beta \left[\Delta_\theta^{b,l}(p, r) - \Delta_{\text{ref}}^{b,l}(p, r) \right]. \tag{2}$$

The MPO loss at stage l is

$$\mathcal{L}_{\text{MPO}}^l = -\frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \frac{1}{|\mathcal{P}_b^l||\mathcal{R}_b^l|} \sum_{p \in \mathcal{P}_b^l} \sum_{r \in \mathcal{R}_b^l} \log \sigma \Big(\hat{r}_\theta^{b,l}(p) - \hat{r}_\theta^{b,l}(r) \Big). \tag{3}$$

Here, $\mathcal{B}$ denotes the set of training samples in the current batch, and we follow the original DPO setting (Rafailov et al., 2023) by setting $\beta = 0.1$. The preference annotations identify which prototypes describe the image but do not specify their relative importance. The loss therefore ranks every preferred expert above the rejected set without ordering the preferred experts, avoiding unsupported competition among coexisting attributes. Thus, P3 is a single-label special case, while P4 and P5 preserve multiple coexisting preferences. The overall routing loss is $\mathcal{L}_{\text{MPO}} = \sum_{l \in \{3,4,5\}} \mathcal{L}_{\text{MPO}}^l$.

3.4 Routing Execution and Training

The learned routing distribution is converted into sparse expert execution, while the complete detector is optimized through a stage-dependent objective.

Expert activation and feature aggregation. Each expert is a lightweight network comprising a convolutional layer followed by CBAM (Woo et al., 2018); architectural details are provided in Appendix C. The learned policy distribution determines which experts participate in feature computation. Let K_l denote the number of activated experts at stage l, with $(K_3, K_4, K_5) = (1, 2, 3)$; the notation K below refers to the corresponding stage-specific value K_l. For sample b at stage l, the selected expert indices are $\mathcal{I}_{b,K}^l = \text{TopK}\left(\pi_\theta^l(\cdot \mid \mathbf{F}_b^l)\right)$. The selected probabilities are renormalized as

$$\hat{w}_{b,i}^l = \frac{\pi_\theta^l(i \mid \mathbf{F}_b^l)}{\sum_{j \in \mathcal{I}_{b,K}^l} \pi_\theta^l(j \mid \mathbf{F}_b^l)}, \qquad i \in \mathcal{I}_{b,K}^l. \tag{4}$$

Only the selected experts are evaluated, and their outputs are aggregated according to the routing weights:

$$\mathbf{Y}_{\text{routed},b}^l = \sum_{i \in \mathcal{I}_{b,K}^l} \hat{w}_{b,i}^l \, \text{Expert}_i^l(\mathbf{F}_b^l). \tag{5}$$

For **shared expert aggregation**, let $\text{Expert}_s^l(\cdot)$ denote the shared expert at stage l. A shared expert is evaluated for every sample to provide a common feature pathway. The final output is the sum of the routed and shared expert features, $\mathbf{Y}_{\text{out},b}^l = \mathbf{Y}_{\text{routed},b}^l + \text{Expert}_s^l(\mathbf{F}_b^l)$.

Training objectives. We use a stage-dependent training objective and apply the routing z-loss (Zoph et al., 2022) to regularize the scale of the policy logits. At stage l, it is defined as $\mathcal{L}_z^l = \frac{1}{B} \sum_{b=1}^B \left(\log \sum_{e=1}^{E_l} \exp(z_{b,e}^l) \right)^2$, and $\mathcal{L}_z = \sum_{l \in \{3,4,5\}} \mathcal{L}_z^l$. Let t denote the current training epoch. The overall objective is

$$\mathcal{L}_{\text{total}}^{(t)} = \begin{cases} \mathcal{L}_{\text{YOLO}} + \lambda_z \mathcal{L}_z, & 1 \le t \le T_c, \\ \mathcal{L}_{\text{YOLO}} + \lambda_{\text{MPO}} \mathcal{L}_{\text{MPO}} + \lambda_z \mathcal{L}_z, & t > T_c, \end{cases} \tag{6}$$

where $\mathcal{L}_{\text{YOLO}}$ combines the standard classification and regression losses, while λ_{MPO} weights preference supervision. Following ST-MoE (Zoph et al., 2022), we set $\lambda_z = 0.001$. The complete training procedure is summarized in Appendix D.

Table 1: Performance comparison across different weather conditions. Bold represents the best, and underline represents the second. Our entries reported as mean±standard deviation are computed over three independent runs with random seeds 42, 43, and 44.

Method	Backbone	Reference	mAP					
			Day Clear	Night Clear	Dusk Rainy	Night Rainy	Day Foggy	Average
Faster R-CNN (Ren et al., 2017)	ResNet101	TPAMI'2017	48.1	34.4	26.0	12.4	32.0	26.2
IBN-Net (Pan et al., 2018)	ResNet101	ECCV'2018	49.7	32.1	26.1	14.3	29.6	25.5
SW (Pan et al., 2019)	ResNet101	ICCV'2019	50.6	33.4	26.3	13.7	30.8	26.1
IterNorm (Huang et al., 2019)	ResNet101	CVPR'2019	43.9	29.6	22.8	12.6	28.4	23.4
ISW (Choi et al., 2021)	ResNet101	CVPR'2021	51.3	33.2	25.9	14.1	31.8	26.3
S-DGOD (Wu & Deng, 2022)	ResNet101	CVPR'2022	56.1	36.6	28.2	16.6	33.5	28.7
CLIPGap (Vidit et al., 2023)	ResNet101	CVPR'2023	51.3	36.9	32.3	18.7	38.5	31.6
Prompt-D (Li et al., 2024)	ResNet101	CVPR'2024	53.6	38.5	33.7	19.2	39.1	32.6
G-NAS (Wu et al., 2024)	ResNet101	AAAI'2024	58.4	45.0	35.1	17.4	36.4	33.5
UFR (Liu et al., 2024)	ResNet101	CVPR'2024	58.6	40.8	33.2	19.2	39.6	33.2
PhysAug (Xu et al., 2025)	ResNet101	AAAI'2025	60.2	44.9	41.2	23.1	40.8	37.5
SDG-DiffDet (Jiang et al., 2025)	ResNet101	ICCV'2025	61.8	44.7	40.9	26.6	<u>43.2</u>	38.9
FWCL (Guo et al., 2025)	ResNet101	TMM'2025	55.5	37.5	32.6	18.9	32.3	30.3
SECoT (Zhang et al., 2025)	ResNet101	CVPR'2025	55.4	42.0	39.2	24.5	40.6	36.6
F-KAN (Zhang et al., 2026)	ResNet101	TIP'2026	59.6	46.4	39.2	24.7	41.9	38.1
MoPE (ours)	ResNet101	-	59.67±0.15	47.67±0.25	<u>44.53±0.46</u>	<u>29.73±0.15</u>	41.63±0.23	<u>40.89±0.04</u>
YOLO-ESMoE (Lin et al., 2026)	CSPNet-L	CVPR'2026	57.53±0.06	41.23±0.35	27.27±0.72	12.17±1.08	35.33±0.35	29.00±0.49
F-KAN (Zhang et al., 2026)	CSPNet-L	TIP'2026	<u>62.6</u>	<u>49.2</u>	40.0	25.1	41.4	38.9
MoPE (ours)	CSPNet-L	-	**62.77±0.21**	**51.30±0.53**	**47.87±0.47**	**32.70±0.66**	**44.97±0.12**	**44.21±0.38**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. Diverse Weather Dataset (Wu & Deng, 2022) comprises five weather and illumination conditions: Daytime-Clear, Daytime-Foggy, Dusk-Rainy, Night-Clear, and Night-Rainy. We use Daytime-Clear as the source domain, with 19,395 training and 8,313 validation images, and evaluate on the four remaining unseen domains, containing 3,775, 3,501, 26,158, and 2,494 images, respectively. All subsets contain annotations for seven categories: person, car, bike, rider, motor, bus, and truck. For **Real-to-Artistic Dataset**, we use Pascal VOC as the sole source domain and Clipart1k, Watercolor2k, and Comic2k as unseen artistic domains. Pascal VOC contains 16,551 training and 4,952 validation images, while the three target datasets contain 1,000, 2,000, and 2,000 test images, respectively. Pascal VOC and Clipart1k cover 20 categories, whereas Watercolor2k and Comic2k contain six-category subsets of the VOC label space.

Implementation details. We adopt YOLO-ESMoE (Lin et al., 2026) as the baseline framework. To evaluate its generality, we employ CSPNet and ResNet101 as the backbone networks. All models use weights pretrained on MS COCO 2017 (Lin et al., 2014) and 640×640 input images. We train them for 80 epochs using SGD with an initial learning rate of 0.01. We set $T_c = 30$ and $\lambda_{\text{MPO}} = 1$. We report mAP@0.5 for detection accuracy and use the number of parameters and GFLOPs to measure model complexity. Additional implementation details are provided in Appendix C.

4.2 COMPARISON WITH STATE-OF-THE-ARTS

We compare MoPE with representative SDGOD approaches, including S-DGOD (Wu & Deng, 2022), CLIPGap (Vidit et al., 2023), UFR (Liu et al., 2024), FWCL (Guo et al., 2025), SDG-DiffDet (Jiang et al., 2025), SECoT (Zhang et al., 2025), and F-KAN (Zhang et al., 2026). Tables 1, 2, and 3 report the generalization performance and model complexity, with detailed per-class analyses provided in Appendix H. These comparisons lead to three main observations.

(1) Strong generalization across diverse weather and artistic shifts. On the Diverse Weather dataset with ResNet-101, MoPE reaches an average mAP of **40.9%**, exceeding the strongest prior result in the table, SDG-DiffDet, by **2.0%** and achieving the highest performance on Night-Clear, Dusk-Rainy, and Night-Rainy. With CSPNet-L, MoPE attains **44.2%** average mAP, outperforming F-KAN by **5.3%** and YOLO-ESMoE by **15.2%**. The gains over F-KAN are particularly pronounced on Dusk-Rainy (+**7.9%**) and Night-Rainy (+**7.6%**). On the Real-to-Artistic dataset, MoPE achieves **54.4%** average mAP, improving over F-KAN by **6.3%** and YOLO-ESMoE by **22.6%**. It yields substantial gains on Clipart (+**12.4%**) and Comic (+**6.5%**) while remaining comparable on Watercolor (61.83% versus 61.9%). These results demonstrate consistent generalization across severe illumination, weather, and appearance shifts.

Table 2: Performance comparison across different artistic conditions.

Method	Backbone	Reference	mAP				
			VOC	Clipart	Watercolor	Comic	Average
Faster R-CNN (Ren et al., 2017)	ResNet101	TPAMI'2017	82.4	25.7	44.5	18.9	29.7
CLIPGap (Vidit et al., 2023)	ResNet101	CVPR'2023	80.5	36.7	50.7	29.4	38.9
NP (Fan et al., 2023)	ResNet101	ICLR'2023	79.2	35.4	53.3	28.9	39.2
FWCL (Guo et al., 2025)	ResNet101	TMM'2025	80.6	36.6	53.9	35.2	41.9
AFDA (Danish et al., 2024)	ResNet101	CVPR'2024	80.1	38.9	57.4	33.2	43.2
SECoT (Zhang et al., 2025)	ResNet101	CVPR'2025	82.9	40.2	57.5	34.8	44.2
SDG-DiffDet (Jiang et al., 2025)	ResNet101	ICCV'2025	86.2	42.1	60.9	36.3	46.4
F-KAN (Zhang et al., 2026)	ResNet101	TIP'2026	84.7	42.0	58.3	36.4	45.6
MoPE (ours)	ResNet101	-	88.50±0.20	54.10±0.44	60.03±0.93	44.40±0.52	52.84±0.53
YOLO-ESMoE (Lin et al., 2026)	CSPNet-L	CVPR'2026	83.13±0.15	32.03±0.93	45.90±0.96	17.50±0.78	31.81±0.41
F-KAN (Zhang et al., 2026)	CSPNet-L	TIP'2026	89.1	43.8	**61.9**	38.7	48.1
MoPE (ours)	CSPNet-L	-	**89.90±0.36**	**56.17±0.84**	61.83±0.47	**45.17±1.04**	**54.39±0.44**

Table 3: Model complexity and average mAP on four unseen domains.

ResNet101 Backbone				CSPNet Backbones			
Method	Params	GFLOPs	mAP	Method	Params	GFLOPs	mAP
Faster R-CNN (Ren et al., 2017)	**90.8**	267.9	26.2	F-KAN (S) (Zhang et al., 2026)	8.1	28.0	30.1
S-DGOD (Wu & Deng, 2022)	120.4	284.0	28.7	MoPE (S)	26.7	29.0	32.3
CLIPGap (Vidit et al., 2023)	190.2	423.0	31.6	F-KAN (M) (Zhang et al., 2026)	23.4	79.0	32.4
UFR (Liu et al., 2024)	184.2	323.0	33.2	MoPE (M)	49.4	90.5	40.6
F-KAN (Zhang et al., 2026)	140.6	264.0	38.1	F-KAN (L) (Zhang et al., 2026)	**30.8**	134.0	38.9
MoPE (ours)	92.4	**214.4**	**40.9**	MoPE (L)	55.6	**112.1**	**44.2**

(2) Improved target-domain generalization without sacrificing source-domain accuracy. Relative to YOLO-ESMoE under the same CSPNet-L backbone, MoPE improves source-domain mAP from 57.53% to 62.77% on Day-Clear and from 83.13% to 89.90% on Pascal VOC. It also slightly exceeds F-KAN on both source domains (62.77% versus 62.6% and 89.90% versus 89.1%). The simultaneous improvements on source and unseen domains indicate that MoPE does not obtain cross-domain robustness by trading off source-domain discrimination.

(3) Favorable trade-off between accuracy and computational efficiency. With ResNet-101, MoPE achieves higher average detection accuracy than F-KAN while using substantially fewer parameters and less computation, demonstrating that its performance gain does not rely on increased model complexity. Across the CSPNet variants, MoPE contains more total parameters because MoE maintains multiple expert branches. Nevertheless, sparse routing evaluates only the selected prototype experts for each input, keeping the executed computation broadly comparable while consistently improving detection accuracy. MoPE therefore benefits from a richer set of specialized experts without a proportional increase in inference computation.

4.3 Ablation Studies and Discussions

We conduct ablation studies on four unseen domains of the Diverse Weather Dataset, evaluating the contributions of MSA and MPO routing and analyzing the routing selections of the trained model. Additional analyses of routing hyperparameters, routing collapse, cross-architecture generalization, and qualitative results are provided in Appendices E, E.3, E.4, and H.8, respectively.

Effectiveness of each component. Table 4 isolates the effects of MPO routing and MSA. MPO alone raises the average target-domain mAP from 29.00 to 35.97 because prototype preferences directly constrain the relative routing probabilities of preferred and rejected experts, teaching the router to prioritize experts that match the scene attributes of each input. MSA achieves 40.30 mAP by introducing broader weather and appearance variations while preserving object geometry. More importantly, these variations enrich the combinations of prototype preference labels, providing MPO with more diverse input-preference correspondences and further strengthening expert specialization. Consequently, combining MSA and MPO reaches 44.21 mAP and improves every target domain. This synergy arises because MSA expands both visual diversity and preference supervision, while MPO converts them into input-adaptive expert routing. The source-domain improvement from 57.53 to 62.77 further shows that the generalization gains do not compromise in-domain detection.

Routing selection analysis. To determine whether prototype preference supervision prevents routing collapse and produces input-dependent expert allocation, we compute each expert's share of all

Table 4: Component ablation of MoPE.

Components			Source Domain	Target Domains				
Baseline	MPO	Routing	MSA	Day Clear	Night Clear	Dusk Rainy	Night Rainy	Day Foggy	Average
✓	×	×		57.53±0.06	41.23±0.35	27.27±0.72	12.17±1.08	35.33±0.35	29.00±0.49
✓	✓	×		62.40±0.10	47.47±0.12	36.67±0.50	18.97±0.93	40.77±0.15	35.97±0.38
✓	×	✓		59.90±0.26	47.63±0.32	42.60±0.46	28.60±0.53	42.37±0.06	40.30±0.04
✓	✓	✓		**62.77±0.21**	**51.30±0.53**	**47.87±0.47**	**32.70±0.66**	**44.97±0.12**	**44.21±0.38**

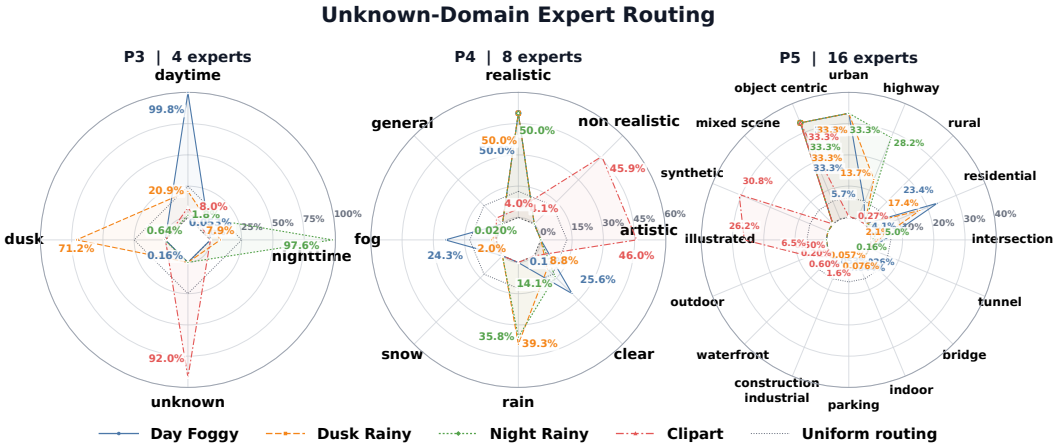

Figure 3: Expert activation distributions across unseen domains during inference. *Zoom in for best view.*

top-K routing slots. As shown in Fig. 3, we visualize four representative unseen domains covering illumination, weather, and artistic shifts. At P3, Day Foggy and Night Rainy allocate 99.8% and 97.6% of their routing slots to the daytime and nighttime experts, respectively. Dusk Rainy predominantly activates the dusk expert at 71.2%, while 20.9% of its routing slots are assigned to the daytime expert. This mixed response is consistent with the gradual illumination transition at dusk, as some scenes retain visual characteristics similar to late daytime. Clipart instead concentrates on the unknown expert because stylized images often lack reliable temporal cues. At P4, the two rainy domains emphasize the rain expert. Day Foggy distributes comparable activation shares to the fog and clear experts, reflecting variations in fog density: scenes with light fog may preserve visibility and contrast similar to clear weather. Clipart primarily activates the artistic and non-realistic experts, consistent with its graphical appearance. At P5, road-scene domains compose urban-street, highway or residential-street, and general object-centric experts, whereas Clipart favors the graphic-synthetic and illustrated-scene experts. These results show that the router captures dominant scene attributes while accommodating gradual or ambiguous visual transitions, thereby producing semantically meaningful and input-dependent expert compositions.

5 Conclusion

We proposed a Mixture of Prototype Experts framework for single domain generalized object detection (MoPE). MoPE describes visual conditions using predefined atomic prototypes along the photometric, textural, and spatial dimensions. It uses multi-label prototype preferences to optimize the router through MPO, encouraging the router to assign higher probabilities to semantically relevant prototype experts. The model then dynamically activates and aggregates the selected prototype experts together with a shared expert for each input. MSA further broadens the visual variations in the source training data without using any target-domain images or statistics. Extensive experiments on diverse weather and artistic benchmarks demonstrate that MoPE consistently improves generalization across unseen domains and different detector backbones while maintaining a favorable trade-off between detection accuracy and computational efficiency. These results validate multi-label preference-guided expert composition as an effective approach to adaptive feature representation under complex domain shifts.

AI USE STATEMENT

Generative AI tools were used in both the research workflow and manuscript preparation. Specifically, FLUX.2 was employed for offline multi-style image generation as part of MSA (Sec. 3.2.2 and Appendix G), and GPT-5.6-sol was used to construct image-level prototype preference annotations (Sec. 3.3.1). Generative AI tools also assisted with manuscript language editing, figure refinement, code debugging, and the interpretation and presentation of experimental results. AI-assisted manuscript content and code were reviewed by the authors, and the generated images were assessed through the sampling-based quality-control procedure described in Appendix G. All reported experimental results and scientific claims were independently checked by the authors, who take full responsibility for the final content of the paper.

REFERENCES

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

Yuehai Chen, Qingzhong Wang, Jing Yang, Badong Chen, Haoyi Xiong, and Shaoyi Du. Cscc: Cross-scene crowd counting via learning to diversify for domain generalization. *IEEE Transactions on Multimedia*, 27:3320–3330, 2025. doi: 10.1109/TMM.2025.3535302.

Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T. Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11580–11590, June 2021.

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 (CVPR)*, pp. 17732–17742, June 2024.

Qi Fan, Mattia Segu, Yu-Wing Tai, Fisher Yu, Chi-Keung Tang, Bernt Schiele, and Dengxin Dai. Towards robust object detection invariant to real-world domain shifts. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*. OpenReview, 2023.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Jacob Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

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*, 2025.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, 2024.

Jian Hu, Shaogang Gong, Weitong Cai, and Junchi Yan. Class-aware diversified augmentation for open-set single domain generalization. *IEEE Transactions on Multimedia*, 28:1779–1789, 2026a. doi: 10.1109/TMM.2025.3645625.

Yingxiang Hu, Yanbo Liu, Guo Cao, and Jin Wang. Crowd counting via single-domain generalization. *IEEE Transactions on Multimedia*, pp. 1–16, 2026b. doi: 10.1109/TMM.2026.3668677.

Lei Huang, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4874–4883, 2019.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

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.

Pengkun Jiao, Na Zhao, Jingjing Chen, and Yu-Gang Jiang. Domain expansion and boundary growth for open-set single-source domain generalization. *IEEE Transactions on Multimedia*, 27:3925–3938, 2025. doi: 10.1109/TMM.2025.3535334.

Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Style normalization and restitution for domain generalization and adaptation. *IEEE Transactions on Multimedia*, 24:3636–3651, 2022. doi: 10.1109/TMM.2021.3104379.

Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. URL `https://github.com/ultralytics/ultralytics`.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. URL `https://github.com/ultralytics/ultralytics`.

Wooju Lee, Dasol Hong, Hyungtae Lim, and Hyun Myung. Object-aware domain generalization for object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):2947–2955, Mar. 2024. doi: 10.1609/aaai.v38i4.28076. URL `https://ojs.aaai.org/index.php/AAAI/article/view/28076`.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Z. Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *ArXiv*, abs/2006.16668, 2020. URL `https://api.semanticscholar.org/CorpusID:220265858`.

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.

Xinhui Li, Mingjia Li, Xiaopeng Li, and Xiaojie Guo. Learning generalized knowledge from a single domain on urban-scene segmentation. *IEEE Transactions on Multimedia*, 25:7635–7646, 2023. doi: 10.1109/TMM.2022.3224328.

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 (ICCV)*, pp. 8771–8780, October 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Xu Lin, Jinlong Peng, Zhenye Gan, Jiawen Zhu, and Jun Liu. YOLO-Master: Moe-accelerated with specialized transformers for enhanced real-time detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the european conference on computer vision (ECCV)*, pp. 464–479, 2018.

Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1863–1871, 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *ArXiv*, abs/1701.06538, 2017. URL https://api.semanticscholar.org/CorpusID:12462234.

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.

Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 834–843, October 2021.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

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.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. In *International Conference on Machine Learning*, 2024. URL https://api.semanticscholar.org/CorpusID:267028540.

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 (CVPR)*, pp. 8103–8112, June 2023.

Xiaoran Xu, Jiangang Yang, Wenhui Shi, Siyuan Ding, Luqing Luo, and Jian Liu. Physaug: A physical-guided and frequency-based data augmentation for single-domain generalized object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21815–21823, 2025.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pp. 818–833. Springer, 2014. doi: 10.1007/978-3-319-10590-1_53.

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. Fourier-kan: Feature distribution decomposition and recombination for unknown-domain object detection. *IEEE Transactions on Image Processing*, 2026.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.

A Supplementary Material Overview

This supplementary material is organized as follows. Appendix B reviews preference optimization. Appendix C provides implementation details, and Appendix D presents the complete MoPE training algorithm. Appendix E reports routing ablations, analyzes routing collapse without explicit preference supervision, and presents cross-architecture experiments. Appendix F provides the complete prototype taxonomy, while Appendix G presents the MSA prompt bank and analyzes the consistency and visual quality of the generated images. Appendix H presents detailed per-class results and additional visualizations, while Appendix I discusses limitations and future work.

B Background on Preference Optimization

Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) aligns language models with human preferences by learning a reward model from pairwise preferences and optimizing the policy through reinforcement learning. DPO (Rafailov et al., 2023) reparameterizes the KL-constrained RLHF objective using log-probability ratios between the policy and a reference model, enabling direct policy learning from preference pairs without explicit reward modeling or reinforcement learning. CPO (Xu et al., 2024) combines reference-free contrastive preference learning with supervised likelihood maximization, whereas ORPO (Hong et al., 2024) integrates supervised fine-tuning and odds-ratio optimization into a single reference-free objective. The reference-based preference optimization formulation provides a natural way to supervise expert routing using semantic prototype preferences.

C Additional Implementation Details

Computational setup. All experiments are conducted on 4 NVIDIA H200 GPUs with a total batch size of 64. Peak GPU memory usage during training is approximately 20 GB per GPU. Training takes approximately 3 minutes per epoch and 4 hours in total.

Prototype expert architecture. We instantiate 4, 8, and 16 prototype experts at P3, P4, and P5, respectively, yielding 28 prototype experts in total, each independently parameterized and associated with a fixed semantic prototype. Each prototype expert comprises a convolutional layer followed by the Convolutional Block Attention Module (CBAM) (Woo et al., 2018), rather than a conventional stack of convolutional layers. CBAM sequentially applies channel and spatial attention to recalibrate feature responses along both dimensions. Compared with a multi-layer convolutional expert, this lightweight structure requires fewer parameters and FLOPs, limiting the overhead of deploying multiple experts. Each stage additionally contains a shared expert implemented as a bias-free 1×1 convolution followed by batch normalization, providing a common feature pathway for all inputs. During inference, sparse routing activates only 1, 2, and 3 prototype experts at P3, P4, and P5, respectively, together with the corresponding shared expert.

Offline MSA generation. We implement MSA as an offline preprocessing step: multi-style images are generated before training and stored locally for reuse across subsequent training runs. This offline generation process takes approximately 2.5 hours.

Preference consistency in Mosaic. Standard Mosaic (Bochkovskiy et al., 2020) samples constituent images from the entire training set and may combine samples with conflicting image-level preferences, such as daytime and nighttime, making the preference supervision of the composite image ambiguous. We therefore introduce preference-consistent Mosaic sampling, where the remaining images for each anchor image are drawn only from the same preference-label group. Detection annotations follow the standard Mosaic transformations, while prototype preferences are merged by set union, deduplicated, and truncated according to K_l. This preserves photometric exclusivity at P3 while supporting compatible multi-label preferences at P4 and P5. Mosaic remains active throughout training. We disable other multi-image augmentation methods to avoid unrestricted mixing of conflicting preferences.

Algorithm 1 Overall Training Procedure of MoPE

Require: Source dataset $\mathcal{D}_s$, prompt bank $\mathcal{G}$, total epochs T, number of cold-start epochs T_c, expert sets $\{\mathcal{E}^l\}_{l=3}^5$, activated expert numbers $\{K_l\}_{l=3}^5$, and weights $\lambda_{\text{MPO}}, \lambda_z$
Ensure: Trained detector f_θ

Offline data construction
Generate $\mathcal{D}_{\text{aug}}$ offline using MSA and inherit the source detection annotations
$\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_s \cup \mathcal{D}_{\text{aug}}$
Assign prototype preference sets $\{\mathcal{P}_b^l\}_{l=3}^5$ to the images in $\mathcal{D}_{\text{train}}$
Stage I: Cold start
for $t = 1$ to T_c **do**
  for each batch $\mathcal{B}$ sampled from $\mathcal{D}_{\text{train}}$ using preference-consistent Mosaic **do**
    $(\widehat{\mathbf{y}}, \{\pi_\theta^l\}_{l=3}^5, \mathcal{L}_z) \leftarrow \text{MoPEForward}(\mathcal{B}; f_\theta, \{K_l\}_{l=3}^5)$
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{YOLO}}(\widehat{\mathbf{y}}, \mathbf{y}) + \lambda_z \mathcal{L}_z$
    Update θ by backpropagating $\mathcal{L}$
  end for
end for
$f_{\text{ref}} \leftarrow \text{Freeze}(\text{CopyBackbone}(f_\theta))$
Stage II: Multi-label preference optimization
for $t = T_c + 1$ to T **do**
  for each batch $\mathcal{B}$ sampled from $\mathcal{D}_{\text{train}}$ using preference-consistent Mosaic **do**
    Compute $\{\pi_{\text{ref}}^l\}_{l=3}^5 \leftarrow \text{ReferenceForward}(\mathcal{B}; f_{\text{ref}})$ without gradients
    $(\widehat{\mathbf{y}}, \{\pi_\theta^l\}_{l=3}^5, \mathcal{L}_z) \leftarrow \text{MoPEForward}(\mathcal{B}; f_\theta, \{K_l\}_{l=3}^5)$
    for $l \in \{3, 4, 5\}$ and $b \in \mathcal{B}$ **do**
      $\mathcal{R}_b^l \leftarrow \mathcal{E}^l \setminus \mathcal{P}_b^l$ and $\mathcal{C}_b^l \leftarrow \mathcal{P}_b^l \times \mathcal{R}_b^l$
    end for
    Compute $\mathcal{L}_{\text{MPO}}$ from $\{\pi_\theta^l\}$, $\{\pi_{\text{ref}}^l\}$, and $\{\mathcal{C}_b^l\}$ using Eq. (3)
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{YOLO}}(\widehat{\mathbf{y}}, \mathbf{y}) + \lambda_{\text{MPO}} \mathcal{L}_{\text{MPO}} + \lambda_z \mathcal{L}_z$
    Update θ by backpropagating $\mathcal{L}$ while keeping f_{ref} fixed
  end for
end for
return f_θ

D MoPE Training Algorithm

Algorithm 1 formalizes how MoPE coordinates detection learning with preference-guided routing. Since preference optimization requires a stable and task-aware reference policy, the detector first learns detection features and an initial routing policy during cold start. The resulting backbone, including its prototype experts and policy routers, is then frozen as f_{ref}. In the subsequent stage, MPO compares the current and reference routing distributions and promotes each preferred expert over the rejected experts.

Before training, MSA generates a style-transformed counterpart for each source image offline. Each generated image inherits the YOLO category and bounding-box annotations of its source image, while the multimodal annotation procedure in Section 3.3.1 assigns prototype preferences to both source and generated images. In Algorithm 1, MoPEForward denotes the complete detector forward pass, including policy routing, top-K expert activation, prototype-expert aggregation, and shared-expert fusion. The reference policy is used only to provide routing distributions during training and is discarded at inference, introducing no additional inference cost.

E Supplementary Ablation Studies

E.1 The Impact of Top-K Number

We evaluate different numbers of activated experts at P3, P4, and P5, denoted by (K_3, K_4, K_5). Since the photometric prototypes at P3 are mutually exclusive, we fix $K_3 = 1$ and vary K_4 and K_5. As shown in Table 5, performance does not increase monotonically with the number of activated experts. Too few experts restrict the composition of coexisting visual attributes, whereas excessive ac-

Table 5: Ablation study on top-K.

(K_3, K_4, K_5)	GFLOPs	Source Domain	Target Domains				
		Day Clear	Night Clear	Dusk Rainy	Night Rainy	Day Foggy	Average
(1, 1, 1)	109.6	63.00±0.10	51.37±0.31	46.80±0.72	32.47±0.25	44.63±0.06	43.82±0.24
(1, 2, 2)	111.5	62.43±0.15	50.23±0.15	46.83±0.29	31.87±0.85	44.40±0.26	43.33±0.23
(1, 2, 3)	112.1	62.77±0.21	51.30±0.53	47.87±0.47	32.70±0.66	44.97±0.12	44.21±0.38
(1, 3, 3)	113.2	62.67±0.12	51.23±0.29	47.53±0.40	32.83±0.64	44.73±0.38	44.08±0.38

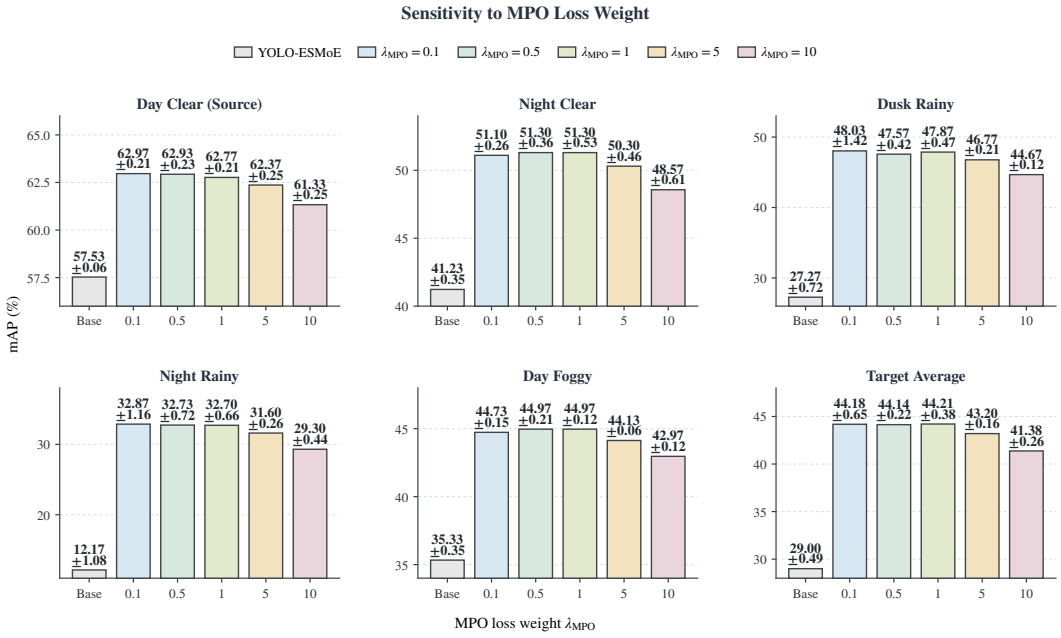

Figure 4: Sensitivity to the MPO loss weight λ_{MPO}. YOLO-ESMoE serves as the baseline. Values above each bar report the mean and standard deviation over three random seeds.

tivation may introduce less relevant expert responses and weaken specialization. The configuration $(1, 2, 3)$ achieves the highest average target-domain mAP of 44.21% with 112.1 GFLOPs. This observation is consistent with prior sparse MoE studies (Fedus et al., 2022; Riquelme et al., 2021): activating a small subset of experts provides a favorable balance between performance and efficiency, while further increasing the number of active experts yields diminishing returns. The asymmetric allocation also follows the prototype hierarchy, as P3 describes mutually exclusive photometric conditions, whereas the finer textural and spatial attributes at P4 and P5 require greater compositional capacity. We therefore adopt $(K_3, K_4, K_5) = (1, 2, 3)$ as the default setting.

E.2 The Impact of the MPO Loss Weight

We assess the sensitivity of MoPE to λ_{MPO} by comparing five weight settings against YOLO-ESMoE. As shown in Fig. 4, MoPE consistently outperforms the baseline across all source and target domains, indicating that its gains do not depend on a narrowly tuned coefficient. In particular, $\lambda_{\text{MPO}} = 0.1$, 0.5, and 1.0 achieve comparable target-domain averages of 44.18%, 44.14%, and 44.21%, respectively. Performance decreases gradually at 5.0 and more noticeably at 10.0, although both settings remain above the baseline. This degradation suggests that excessive preference supervision biases joint optimization toward expert ranking at the expense of classification and localization. We therefore adopt $\lambda_{\text{MPO}} = 1.0$ as the default setting.



Figure 5: Top-2 expert selection shares (%) of YOLO-ESMoE across representative unseen domains (seed 44). Similar activation patterns across domains reveal concentrated and weakly input-adaptive routing.

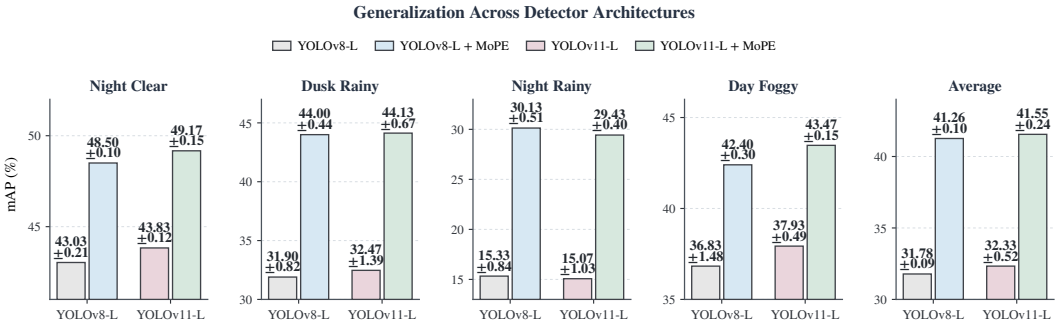

Figure 6: Generalization across detector architectures. Each panel compares the baseline and MoPE variants of YOLOv8-L and YOLOv11-L. Values above the bars report the mean and standard deviation over three random seeds.

E.3 Preference Optimization Mitigates Routing Collapse

To examine routing behavior in the absence of explicit preference supervision, we visualize the Top-2 routing distributions of YOLO-ESMoE across representative weather and artistic domains. Since each image contributes two routing slots, a selection share approaching 50% indicates that the corresponding expert is activated for nearly every input. As shown in Fig. 5, P2 consistently selects Experts 1 and 3, whereas P5 almost exclusively activates Experts 2 and 3 across all domains. P4 similarly favors Experts 1 and 3, with only limited variation in the remaining branch. Although P3 responds to broad visual changes, several distinct domains still converge to the same expert pairs. These patterns show that, without explicit preference supervision, YOLO-ESMoE concentrates routing on a small number of nearly fixed expert combinations, leaving some branches rarely activated and limiting fine-grained adaptation to input variations. In contrast, the MoPE routing distributions in Fig. 3 vary systematically with the photometric, textural, and spatial attributes of individual inputs. This comparison demonstrates that MPO mitigates routing collapse by explicitly increasing the relative probabilities of input-relevant prototype experts, thereby enabling semantically structured and input-adaptive expert composition.

E.4 Generalization Capability of MoPE Across Different Model Architectures

As shown in Fig. 6, we further validate the effectiveness of the proposed method on YOLOv8 (Jocher et al., 2023) and YOLOv11 (Jocher & Qiu, 2024), which are two widely used and representative object detection frameworks. This evaluation provides a more comprehensive assessment of the generality and compatibility of MoPE across different architectures. The results show that MoPE consistently brings significant performance improvements on both models. Under YOLOv8-L, the av-

erage mAP improves from 31.78% to 41.26%, yielding an absolute gain of 9.48%. Similarly, under YOLOv11-L, the average mAP increases from 32.33% to 41.55%, corresponding to an absolute improvement of 9.22%. More importantly, larger gains are observed under challenging scenarios such as *Night-Rainy*, demonstrating the strong adaptability of MoPE to complex unseen domains. These results indicate that MoPE does not rely on a specific detector architecture, but instead serves as a general dynamic expert mechanism that consistently enhances cross-domain generalization across different models.

F Atomic Prototype Taxonomy

The prototype taxonomy is constructed exclusively from common visual concepts observed in the source-domain data and is fixed before target-domain evaluation. Tables 6 and 7 provide the complete expert-index mappings and semantic definitions used for preference annotation and routing supervision.

Table 6: Photometric and textural prototypes at P3 and P4.

Stage	Index	Prototype	Semantic definition
P3	0	*daytime*	Normal daylight between sunrise and sunset.
P3	1	*nighttime*	Nighttime, low illumination, or scenes dominated by artificial light.
P3	2	*unknown*	The time of day cannot be determined reliably, such as in indoor or non-realistic scenes.
P3	3	*dusk*	Transitional illumination around sunset, evening, or twilight.
P4	0	*realistic*	Photographic appearance of real-world scenes.
P4	1	*non-realistic*	Non-photographic graphical appearance, such as clipart, icons, vectors, or synthetic renderings.
P4	2	*artistic*	Artistic appearance, such as watercolor, painting, sketch, comic, manga, or anime.
P4	3	*clear*	Clear, dry weather with normal visibility.
P4	4	*rain*	Rainfall, rain streaks, wet roads, or other rainy appearances.
P4	5	*snow*	Snowfall, accumulated snow, ice, or winter precipitation.
P4	6	*fog*	Fog, mist, or haze that reduces visibility.
P4	7	*general*	General appearance for mixed, unknown, or otherwise broadly applicable visual conditions.

G Prompt Bank and Consistency Analysis of MSA

Prompt bank. MSA uses a fixed prompt bank derived from generic photometric and textural attributes. It comprises six weather–illumination transformations and two broad appearance transformations. Before offline generation, we select a set of prompts from this fixed bank and randomly assign the selected prompts to the source images. Each image is then translated once using FLUX.2. The prompts specify the desired appearance changes while requiring the original composition, objects, and overall scene details to remain unchanged, allowing each generated image to inherit the detection annotations of its source image. The prompt bank is fixed before training and is constructed without using target-domain images, annotations, or statistics. Table 8 presents the complete prompt bank.

Annotation consistency and visual quality. Since MSA directly inherits the object annotations of each source image, reliable annotation transfer requires the generation process to preserve scene content and object geometry. We therefore randomly select paired source and generated images and overlay the original YOLO bounding boxes on both images for qualitative comparison. As shown in Fig. 7, MSA produces diverse changes in weather, illumination, and visual appearance while preserving object categories, locations, scales, viewpoints, and overall scene layouts. The inherited bounding boxes remain well aligned with the corresponding objects in the generated images, without obvious target omission or geometric distortion. These observations demonstrate that the generated images maintain reliable visual quality and annotation consistency, supporting the direct reuse of the source detection annotations. Beyond the examples shown, we manually inspected a substan-

Table 7: Spatial prototypes describing scene layout at P5.

Index	Prototype	Semantic definition
0	*urban street*	Dense urban or commercial streets with buildings, vehicles, pedestrians, and sidewalks.
1	*highway*	Highways, expressways, or multi-lane arterial roads with long road perspective.
2	*rural road*	Rural, mountain, or suburban roads surrounded by open land, vegetation, or sparse buildings.
3	*residential street*	Community or residential roads with houses, parked vehicles, and local branches.
4	*intersection*	Intersections, roundabouts, crossings, traffic lights, or merging traffic flows.
5	*tunnel/underpass*	Tunnels, underpasses, culverts, underground parking areas, or enclosed road spaces.
6	*bridge/elevated*	Bridges, elevated roads, flyovers, or multi-level road structures.
7	*indoor public*	Stations, airports, malls, exhibition halls, corridors, or other indoor public spaces.
8	*parking/service*	Parking lots, gas stations, service areas, toll stations, or maintenance areas.
9	*construction/industrial*	Construction sites, industrial parks, warehouses, factories, or engineering facilities.
10	*waterfront/open*	Waterfront roads, docks, coasts, riverbanks, lakesides, or large open public spaces.
11	*natural outdoor*	Forests, grasslands, mountains, wilderness, parks, or other natural outdoor scenes.
12	*illustrated scene*	Paintings, comics, anime, watercolor, or other illustrated scenes with explicit spatial structure.
13	*graphic synthetic scene*	Clipart, vectors, icons, game imagery, 3D renderings, or other synthetic graphical scenes.
14	*mixed complex scene*	Scenes combining multiple spatial structures, such as streets, bridges, elevated roads, or construction areas.
15	*general object-centric*	Images dominated by local objects or close-up subjects without a clearly identifiable global scene structure.

Table 8: Complete prompt bank for offline multi-style image generation.

Generation prompt
Change the weather in the scene from a clear daytime to a rainy night, with rain slanting on the car window, in a city style, while keeping all other content unchanged.
Change the weather in the scene from a clear daytime to a foggy daytime, with fog spreading on the car windows, while keeping everything else unchanged.
Change the weather in the scene of the image from a clear daytime to a rainy afternoon, with rain slanting on the car windows, in a city style, and keep everything else unchanged.
Change the weather of the scene in the picture from a clear daytime to a clear nighttime, with a pitch-dark sky and an urban style, while keeping all other elements unchanged.
Change the weather in the scene from a clear daytime to a snowy daytime, with falling snow and snow-covered surfaces, while keeping all other content unchanged.
Change the weather in the scene from a clear daytime to a snowy night, with falling snow and snow-covered surfaces, while keeping all other content unchanged.
Convert this realistic photograph into an artistic style. Preserve the original composition, objects, lighting, and overall scene details.
Convert this realistic photograph into a non-realistic style. Preserve the original composition, objects, lighting, and overall scene details.

tially larger random subset for object retention and inherited-box alignment, and observed consistent content and annotation quality. Only representative cases are shown due to space constraints; the complete generated dataset will be released for independent inspection and reproducibility.



Figure 7: Qualitative consistency and quality analysis of MSA. The original YOLO bounding boxes are overlaid on randomly selected source images and their corresponding generated images. MSA changes weather, illumination, and visual appearance while preserving object content and scene geometry.

Table 9: Per-Class Results (%) on the DayClear to NightClear.

Method	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
YOLO-ESMoE (Lin et al., 2026)	44.1	32.9	68.2	14.5	50.9	29.8	46.0	40.9
IBN-Net (Pan et al., 2018)	37.8	27.3	49.6	15.1	29.2	27.1	38.9	32.1
SW (Pan et al., 2019)	38.7	29.2	49.8	16.6	31.5	28.0	40.2	33.4
IterNorm (Huang et al., 2019)	38.5	23.5	38.9	15.8	26.6	25.9	38.1	29.6
ISW (Choi et al., 2021)	38.5	28.5	49.6	15.4	31.9	27.5	41.3	33.2
S-DGOD (Wu & Deng, 2022)	40.6	35.1	50.7	19.7	34.7	32.1	43.4	36.6
CLIPGap (Vidit et al., 2023)	37.7	34.3	58.0	19.2	37.6	28.5	42.9	36.9
Prompt-D (Li et al., 2024)	40.9	35.0	59.0	21.3	40.4	29.9	42.9	38.5
G-NAS (Wu et al., 2024)	46.9	40.5	67.5	26.5	50.7	35.4	47.8	45.0
UFR (Liu et al., 2024)	43.6	38.1	66.1	14.7	49.1	26.4	47.5	40.8
FWCL (Guo et al., 2025)	39.6	36.4	58.9	18.0	42.6	26.2	40.5	37.5
SECoT (Zhang et al., 2025)	42.0	37.2	66.2	21.6	49.6	33.3	44.2	42.0
F-KAN (Zhang et al., 2026)	50.6	45.3	71.6	25.2	55.5	**41.9**	54.6	49.2
MoPE (ours)	**55.0**	**47.2**	**72.0**	**27.1**	**57.0**	41.1	**55.3**	**50.7**

H Additional Experimental Analysis

Unless otherwise specified, all per-class results in this section are reported from the run with random seed 44, whereas the aggregate mean±standard deviation results in the main text are computed over three independent runs with random seeds 42, 43, and 44.

H.1 Detailed Results on Day Clear to Night Clear

The per-class results are reported in Table 9. MoPE achieves the highest overall mAP of **50.7%**, outperforming F-KAN (Zhang et al., 2026) by **1.5%** and YOLO-ESMoE (Lin et al., 2026) by **9.8%**. Moreover, MoPE achieves the best performance in six of the seven object categories and ranks second on *Rider*. These broad per-class improvements indicate that the prototype experts effectively capture complementary photometric, textural, and spatial patterns. Their input-adaptive composition enables the detector to maintain discriminative features under substantial illumination changes, resulting in robust detection performance in nighttime scenes.

H.2 Detailed Results on Day Clear to Dusk Rainy

Table 10 reports the per-class results under the combined shifts in illumination and weather. MoPE achieves the highest overall mAP of **47.5%**, outperforming F-KAN (Zhang et al., 2026) by **7.5%**

Table 10: Per-Class Results (%) on the DayClear to DuskRainy.

Method	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
YOLO-ESMoE (Lin et al., 2026)	32.1	11.1	67.7	3.1	24.8	8.8	40.0	26.8
IBN-Net (Pan et al., 2018)	37.0	14.8	50.3	11.4	17.3	13.3	38.4	26.1
SW (Pan et al., 2019)	35.2	16.7	50.1	10.4	20.1	13.0	38.8	26.3
IterNorm (Huang et al., 2019)	32.9	14.1	38.9	11.0	15.5	11.6	35.7	22.8
ISW (Choi et al., 2021)	34.7	16.0	50.0	11.1	17.8	12.6	38.8	25.9
S-DGOD (Wu & Deng, 2022)	37.1	19.6	50.9	13.4	19.7	16.3	40.7	28.2
CLIPGap (Vidit et al., 2023)	37.8	22.8	60.7	16.8	26.8	18.7	42.4	32.3
Prompt-D (Li et al., 2024)	39.4	25.2	60.9	20.4	29.9	16.5	43.9	33.7
G-NAS (Wu et al., 2024)	44.6	22.3	66.4	14.7	32.1	19.6	45.8	35.1
UFR (Liu et al., 2024)	37.1	21.8	67.9	16.4	27.4	17.9	43.9	33.2
FWCL (Guo et al., 2025)	37.2	22.6	60.5	17.0	27.0	19.9	43.8	32.6
SECoT (Zhang et al., 2025)	45.8	19.2	74.3	23.4	35.8	22.1	53.5	39.2
F-KAN (Zhang et al., 2026)	48.9	22.5	75.7	23.5	38.3	19.8	56.2	40.0
MoPE (ours)	**55.7**	**34.2**	**78.2**	**28.2**	**47.9**	**28.2**	**60.3**	**47.5**

Table 11: Per-Class Results (%) on the DayClear to NightRainy.

Method	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
YOLO-ESMoE (Lin et al., 2026)	21.6	2.5	31.7	0.1	8.1	2.7	13.5	11.4
IBN-Net (Pan et al., 2018)	24.6	10.0	28.4	0.9	8.3	9.8	18.1	14.3
SW (Pan et al., 2019)	22.3	7.8	27.6	0.2	10.3	10.0	17.7	13.7
IterNorm (Huang et al., 2019)	21.4	6.7	22.0	0.9	9.1	10.6	17.6	12.6
ISW (Choi et al., 2021)	22.5	11.4	26.9	0.4	9.9	9.8	17.5	14.1
S-DGOD (Wu & Deng, 2022)	24.4	11.6	29.5	9.8	10.5	11.4	19.2	16.6
CLIPGap (Vidit et al., 2023)	28.6	12.1	36.1	9.2	12.3	9.6	22.9	18.7
Prompt-D (Li et al., 2024)	25.6	12.1	35.8	10.1	14.2	12.9	22.9	19.2
G-NAS (Wu et al., 2024)	28.6	9.8	38.4	0.1	13.8	9.8	21.4	17.4
UFR (Liu et al., 2024)	29.9	11.8	36.1	9.4	13.1	10.5	23.3	19.2
FWCL (Guo et al., 2025)	28.9	11.2	34.6	9.1	12.1	12.0	23.9	18.9
SECoT (Zhang et al., 2025)	37.9	**17.0**	42.6	**12.9**	16.7	**14.2**	30.6	24.5
F-KAN (Zhang et al., 2026)	42.4	13.2	51.8	9.6	19.7	13.7	38.5	25.1
MoPE (ours)	**52.9**	13.5	**58.8**	9.8	**30.3**	11.9	**47.1**	**32.1**

and YOLO-ESMoE (Lin et al., 2026) by **20.7%**. It ranks first across all seven categories, with particularly clear improvements over F-KAN on *Bike*, *Person*, and *Rider*. Unlike Night Clear, Dusk Rainy simultaneously introduces transitional illumination, reduced visibility, and rain-induced appearance changes. The consistent gains indicate that MoPE can jointly activate complementary photometric and textural experts while preserving spatial information, enabling the detector to accommodate multiple coexisting shifts rather than treating them as a single undifferentiated domain.

H.3 Detailed Results on Day Clear to Night Rainy

Table 11 reports the per-class results under the combined effects of low illumination and rain. MoPE achieves the highest overall mAP of **32.1%**, outperforming F-KAN (Zhang et al., 2026) by **7.0%** and YOLO-ESMoE (Lin et al., 2026) by **20.7%**. It ranks first on *Bus*, *Car*, *Person*, and *Truck*, with particularly clear gains on objects whose appearance is strongly affected by reduced visibility. These improvements arise from the complementary roles of MSA and preference-optimized routing. MSA exposes the detector to diverse combinations of illumination and weather variations during training, reducing its dependence on the clear daytime source distribution. The enriched visual conditions and their prototype preferences further enable the router to combine suitable expert responses for nighttime and rainy inputs. Nevertheless, categories with small or thin structures, such as *Bike*, *Motor*, and *Rider*, remain challenging, indicating that fine-grained visual cues are more susceptible to severe nighttime degradation.

Table 12: Per-Class Results (%) on the DayClear to DayFoggy .

Method	Bus	Bike	Car	Motor	Person	Rider	Truck	mAP
YOLO-ESMoE (Lin et al., 2026)	29.1	27.8	61.3	27.2	40.6	39.3	20.0	35.0
IBN-Net (Pan et al., 2018)	29.9	26.1	44.5	24.4	26.2	33.5	22.4	29.6
SW (Pan et al., 2019)	30.6	26.2	44.6	25.1	30.7	34.6	23.6	30.8
IterNorm (Huang et al., 2019)	29.7	21.8	42.4	24.4	26.0	33.3	21.6	28.4
ISW (Choi et al., 2021)	29.5	26.4	49.2	27.9	30.7	34.8	24.0	31.8
S-DGOD (Wu & Deng, 2022)	32.9	28.0	48.8	29.8	32.5	38.2	24.1	33.5
CLIPGap (Vidit et al., 2023)	36.1	34.3	58.0	33.1	39.0	43.9	25.1	38.5
Prompt-D (Li et al., 2024)	36.1	34.5	58.4	33.3	40.5	44.2	26.2	39.1
G-NAS (Wu et al., 2024)	32.4	31.2	57.7	31.9	38.6	38.5	24.5	36.4
UFR (Liu et al., 2024)	36.9	35.8	61.7	33.7	39.5	42.2	27.5	39.6
FWCL (Guo et al., 2025)	31.4	24.9	50.8	25.7	32.5	37.5	23.5	32.3
SECoT (Zhang et al., 2025)	39.3	36.3	60.8	33.8	39.4	42.7	31.8	40.6
F-KAN (Zhang et al., 2026)	**42.7**	**39.4**	63.0	33.7	45.7	43.2	33.7	41.9
MoPE (ours)	42.2	34.8	**70.2**	**37.5**	**48.2**	**47.7**	**33.9**	**44.9**

Table 13: Per-Class Results (%) on the Pascal VOC to Watercolor2k.

Method	Bicycle	Bird	Car	Cat	Dog	Person	mAP
Faster R-CNN (Ren et al., 2017)	85.7	42.5	36.4	29	18.7	54.5	44.5
YOLO-ESMoE (Lin et al., 2026)	70.7	40.5	55.8	15.5	21.6	64.5	44.8
CLIPGap (Vidit et al., 2023)	86.5	56.2	49.1	40.6	39.6	35.7	50.7
NP (Fan et al., 2023)	87.3	56.2	50.4	42.1	41.8	42.2	53.3
AFDA (Danish et al., 2024)	90.4	51.8	51.9	43.9	35.9	70.2	57.4
SECoT (Zhang et al., 2025)	91.9	56.7	50.4	**48.2**	41.6	56.2	57.5
SDG-DiffDet (Jiang et al., 2025)	**95.2**	56.4	57.2	45.0	44.9	66.6	60.9
F-KAN (Zhang et al., 2026)	92.8	**65.4**	53.3	47.9	**48.9**	63.2	**61.9**
MoPE (ours)	74.7	60.9	**68.5**	42.3	44.4	**76.8**	61.3

H.4 Detailed Results on Day Clear to Day Foggy

Table 12 reports the per-class results under fog-induced visibility degradation. MoPE achieves the highest overall mAP of **44.9%**, outperforming F-KAN (Zhang et al., 2026) by **3.0%** and YOLO-ESMoE (Lin et al., 2026) by **9.9%**. It ranks first in five categories and second on *Bus*, with notable gains over F-KAN on *Car* and *Rider*. Fog reduces image contrast and obscures local appearance cues while largely preserving scene geometry. MSA exposes the detector to varying degrees of fog-induced appearance changes, while preference-optimized routing adaptively combines the corresponding textural experts with spatial experts that retain structural information. Their complementary effects improve detection robustness when object boundaries and textures become less distinguishable.

H.5 Detailed Results on Pascal VOC to Watercolor2k

Table 13 reports the per-class results for generalization from natural images to watercolor-style images. MoPE achieves the second-highest overall mAP of **61.3%**, trailing F-KAN (Zhang et al., 2026) by only **0.6%** while outperforming YOLO-ESMoE (Lin et al., 2026) by **16.5%**. It obtains the best results on *Car* and *Person* and ranks second on *Bird*. Watercolor images replace photographic textures with brush strokes, color diffusion, and simplified object details. MSA reduces reliance on source-domain appearance by exposing the detector to diverse non-realistic and artistic variations, while preference-optimized routing adapts the feature composition to the corresponding textural and spatial patterns. The strong results on *Car* and *Person* indicate effective preservation of object structure across substantial style changes. However, the lower performance on *Bicycle* suggests that thin and repetitive structures remain sensitive to watercolor abstraction.

Table 14: Per-Class Results (%) on the Pascal VOC to Comic2k.

Method	Bicycle	Bird	Car	Cat	Dog	Person	mAP
Faster R-CNN (Ren et al., 2017)	39.7	9.1	23.9	9.1	9.1	22.2	18.9
YOLO-ESMoE (Lin et al., 2026)	24.4	5.2	24.4	3.2	7.2	37.8	17.0
CLIPGap (Vidit et al., 2023)	42.0	15.9	42.0	18.8	24.6	33.3	29.4
NP (Fan et al., 2023)	42.4	18.3	38.8	17.3	24.3	32.2	28.9
AFDA (Danish et al., 2024)	54.1	16.9	30.1	25.0	27.4	45.9	33.2
SECoT (Zhang et al., 2025)	47.0	23.0	45.0	20.9	29.4	43.6	34.8
SDG-DiffDet (Jiang et al., 2025)	**57.6**	24.2	33.2	27.4	29.2	46.4	36.3
F-KAN (Zhang et al., 2026)	56.4	23.4	42.3	30.3	29.7	50.2	38.7
MoPE (ours)	56.4	**29.2**	**56.2**	**32.8**	**33.9**	**64.8**	**45.5**

Table 15: Per-Class Results (%) on the Pascal VOC to Clipart1k.

Method	Plane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	Tv	mAP
Faster R-CNN (Ren et al., 2017)	19.9	51.6	17.0	21.9	27.2	49.6	25.5	9.1	35.1	9.1	25.4	3.0	29.2	48.9	30.1	40.3	9.1	6.7	35.2	21.0	25.7
YOLO-ESMoE (Lin et al., 2026)	11.1	50.6	19.4	18.8	33.7	53.6	27.8	9.4	53.4	14.7	31.9	9.4	24.8	63.3	45.5	44.1	14.0	31.6	35.0	53.4	32.3
CLIPGap (Vidit et al., 2023)	46.1	42.2	41.0	24.9	27.3	43.4	54.6	39.8	33.3	40.1	28.5	33.0	55.3	39.0	45.0	10.7	21.5	37.3	32.7	38.4	36.7
NP (Fan et al., 2023)	40.9	44.3	38.0	24.3	26.8	47.0	56.7	33.8	35.8	31.5	32.4	26.2	45.7	38.8	41.6	10.4	16.0	41.7	38.2	37.9	35.4
AFDA (Danish et al., 2024)	34.4	64.4	22.7	27.0	**45.6**	59.2	32.9	7.0	46.8	55.8	28.9	14.5	44.4	58.0	55.2	52.1	14.8	38.4	42.5	33.9	38.9
SECoT (Zhang et al., 2025)	**51.2**	51.9	**50.2**	28.9	27.0	46.2	55.2	**44.5**	33.6	38.3	34.7	40.2	**59.8**	42.8	49.1	13.1	25.9	34.8	37.9	38.3	40.2
F-KAN (Zhang et al., 2026)	30.1	54.1	49.8	27.0	29.4	52.7	60.4	42.8	55.9	45.5	40.2	**47.0**	49.9	51.3	53.9	24.1	32.0	41.8	42.1	46.8	43.8
MoPE (ours)	48.1	**78.5**	48.6	**49.8**	45.1	**61.6**	**65.1**	25.1	**63.8**	**67.5**	**41.3**	15.9	50.5	**83.2**	**75.3**	**70.3**	**35.2**	**58.9**	**60.9**	**59.5**	**55.2**

H.6 Detailed Results on Pascal VOC to Comic2k

Table 14 reports the per-class results for generalization from natural images to comic-style images. MoPE achieves the highest overall mAP of **45.5%**, outperforming F-KAN (Zhang et al., 2026) by **6.8%** and YOLO-ESMoE (Lin et al., 2026) by **28.5%**. It ranks first in five categories and matches F-KAN on *Bicycle*, with particularly substantial gains on *Car* and *Person*. Comic images replace photographic textures with stylized contours, flat colors, and exaggerated appearance patterns, resulting in a substantial source-to-target discrepancy. MSA introduces diverse non-realistic and artistic observations during training, reducing the detector's dependence on photographic appearance. The associated prototype preferences further guide the router to combine textural and spatial responses suited to illustrated scenes. These complementary effects enable MoPE to preserve object semantics and geometry despite pronounced appearance distortions.

H.7 Detailed Results on Pascal VOC to Clipart1k

Table 15 reports the per-class results for generalization from natural images to clipart images. MoPE achieves the highest overall mAP of **55.2%**, outperforming F-KAN (Zhang et al., 2026) by **11.4%** and YOLO-ESMoE (Lin et al., 2026) by **22.9%**. It ranks first in 14 categories, with particularly substantial gains on *Bicycle*, *Boat*, *Motor*, *Person*, and *Plant*. Clipart images substantially simplify textures, contours, and backgrounds, producing a pronounced discrepancy from photographic source images. The improvements span vehicles, humans, furniture, and plants, indicating that the learned features remain effective across diverse object structures rather than benefiting only a limited set of categories. Together with the results on Watercolor2k and Comic2k, these findings demonstrate consistent robustness across different forms of artistic domain shift. Nevertheless, *Cat* and *Dog* remain challenging because severe shape abstraction and texture removal weaken the fine-grained cues required to distinguish deformable animal categories.

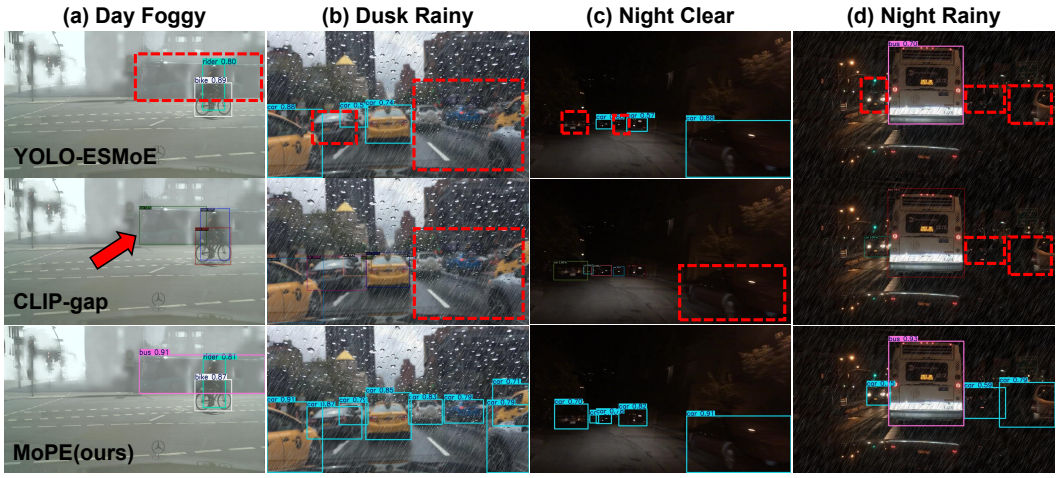

Figure 8: Qualitative detection results under four unseen domains. Red dashed boxes indicate missed detections, while red arrows denote misclassifications. *Zoom in for best view.*

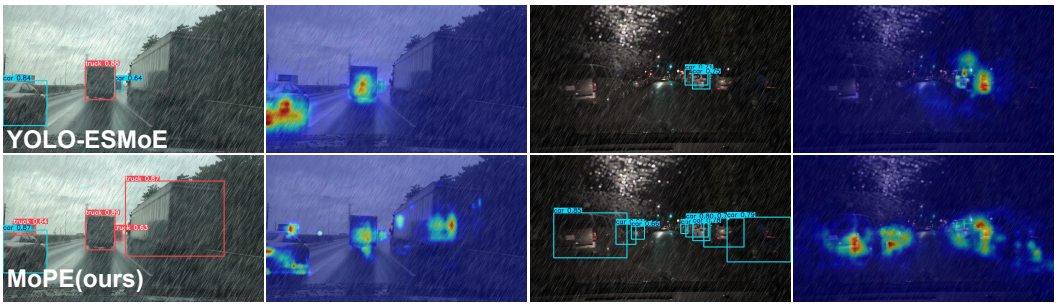

Figure 9: Grad-CAM visualization of feature responses for unseen domains. Brighter regions indicate stronger activation and higher attention to informative object areas.

H.8 Additional Visualization Analysis

Qualitative detection results. Figure 8 compares MoPE with YOLO-ESMoE (Lin et al., 2026) and CLIPGap (Vidit et al., 2023) across four unseen weather domains. Red dashed boxes indicate missed detections, while red arrows indicate misclassifications. Under Daytime-Foggy, YOLO-ESMoE misses several objects, whereas CLIPGap misclassifies a distant *Bus* as a *Car* and produces imprecise bounding boxes. MoPE correctly detects and localizes these targets. Similar improvements are observed under Dusk-Rainy, Night-Clear, and Night-Rainy, where MoPE reduces missed detections, particularly for distant objects and those near image boundaries.

Feature attribution. We further analyze feature responses using Grad-CAM (Gildenblat & contributors, 2021) (Fig. 9). YOLO-ESMoE exhibits sparse and localized activations that fail to cover object regions fully. In contrast, MoPE produces more concentrated and spatially coherent responses that focus on targets while suppressing background interference. This indicates that MoPE learns more structured and discriminative representations through prototype-guided routing and expert aggregation. Moreover, activation patterns align well with detection outcomes: sparse responses correspond to incomplete detections, whereas complete activations lead to accurate predictions. This consistency provides feature-level evidence of the proposed method's effectiveness under unseen domains.

I LIMITATIONS AND FUTURE WORK

Although MPO substantially improves the semantic consistency of expert selection, it provides preference guidance rather than enforcing hard routing constraints. Consequently, inputs near the boundary between visually similar conditions may still be routed to neighboring prototype experts. For example, as shown in Fig. 3, some dusk scenes are assigned to the daytime expert because transitional illumination shares photometric characteristics with late daytime. Such behavior is reasonable under visual ambiguity, yet it also highlights the limitation of representing continuous domain variations using a fixed discrete taxonomy. Moreover, multimodal preference annotation and offline MSA introduce a one-time data-preparation cost, although the resulting annotations and images can be reused across training runs. Future work will explore automatic taxonomy expansion from source-domain observations, more efficient preference construction, and extensions to open-world detection and broader domain shifts.