EviAttr-OW: Evidential Attribute Reasoning for Open World Object Detection

Linhua Ye **Xing Xi** **Ronghua Luo***
School of Computer Science and Engineering
South China University of Technology
Guangzhou, China 510006
cslinhuaye@mail.scut.edu.cn xxyzll@yeah.net rhluo@scut.edu.cn

Abstract

Open World Object Detection (OWOD) is a challenging task that requires detectors to recognize known categories while discovering unlabeled objects and incrementally incorporating them as new categories. Existing methods mainly rely on known-class features to recall unknown objects, while overlooking the semantic pull of known classes in the attribute space and its impact on the interpretability of unknown detection. In this paper, we propose EviAttr-OW, a novel evidential attribute reasoning framework for OWOD that discovers potential unknown objects with interpretable attribute evidence. Specifically, we construct a class-agnostic attribute space that decouples attribute representations from known-class bias and provides a more open attribute description basis for potential unknown objects. We then map candidate-region attribute responses into a Dirichlet evidence distribution, producing known-class predictions and evidential uncertainty to measure the reliability of known-class support. Finally, we derive Known-Class Evidence Deficiency from this evidence distribution and combine it with object evidence, enabling the model to identify unknown objects with evidence-supported objectness and insufficient known-class support. Experiments on the Real-World Object Detection (RWD) benchmark across five real-world application datasets show that EviAttr-OW consistently outperforms existing state-of-the-art (SOTA) methods, achieving +7.3 mAP on unknown classes.

1 Introduction

As a fundamental task in computer vision, object detection has long relied on the closed-set assumption, where the training and test stages share the same category space [1–5]. However, real-world environments are inherently open, with new objects continuously emerging and exhaustive annotation of all possible categories being nearly infeasible. Open World Object Detection (OWOD) [6] has therefore been introduced to enable detectors to discover unlabeled unknown objects while recognizing known categories, and to continuously expand their open-world recognition capability through incremental learning.

Recent OWOD methods mainly mine potential unknown objects along two directions. One line of work is based on feature similarity or objectness assumptions. For example, OW-DETR mines potential objects using attention responses, while PROB estimates probabilistic objectness scores by modeling the feature distribution of known categories. Another line of work leverages semantic knowledge from foundation models for supervision. FOMO, for instance, uses large language models to generate attributes for known categories and provides pseudo labels for unknown objects based on attribute similarity. As shown in the upper part of Fig. 1, such methods retain the attributes most

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

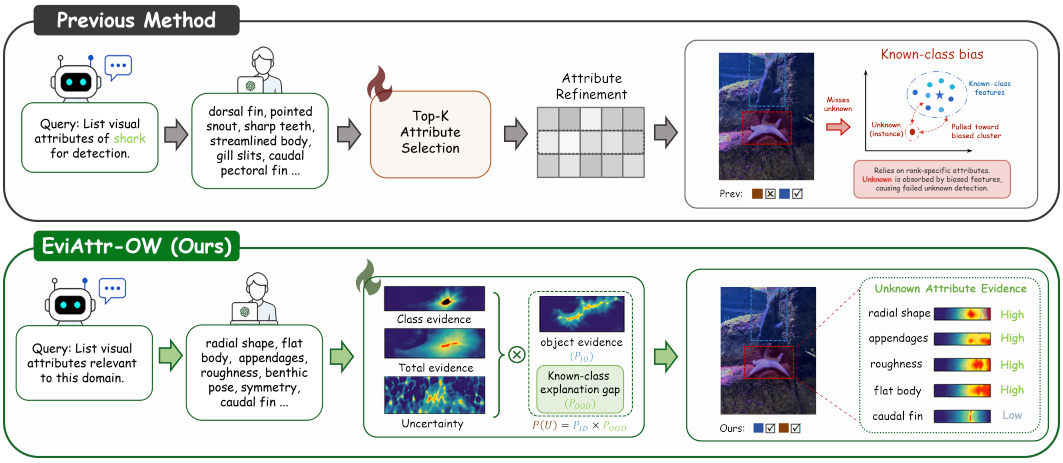

Figure 1: An illustration of EviAttr-OW and previous methods. Previous methods rely on known-category-induced attributes and may absorb unknown objects into biased known-class clusters, while EviAttr-OW discovers unknown objects through class-agnostic attributes and evidential attribute reasoning with interpretable attribute evidence.

relevant to known categories through Top-K attribute selection, making the attribute responses of unknown objects more likely to be pulled toward known-category clusters and thereby forming known-class bias. Therefore, relying solely on attribute similarity makes it difficult to determine whether real unknown objects can be explained by known categories with sufficient evidence, which may cause unknown objects to be absorbed into known-category clusters or misclassified as background.

To address this issue, we propose a novel detection framework, EviAttr-OW, which mines potential unknown objects through evidential attribute reasoning. As shown in the lower part of Fig. 1, EviAttr-OW first constructs a class-agnostic attribute space, where attribute descriptions are no longer driven by the semantics of specific known categories. Then, the visual representations of candidate regions are matched with class-agnostic attributes to obtain fine-grained attribute responses, which are further converted into class-level evidence. Unlike methods that rely only on attribute similarity, this evidence distribution simultaneously characterizes the support of a candidate region for known categories, the total evidence strength, and the predictive uncertainty. Based on this, EviAttr-OW combines object evidence with Known-Class Evidence Deficiency to estimate the unknown probability, grounding unknown decisions in interpretable attribute evidence. In this way, real unknown objects can be identified by their clear object evidence and insufficient known-class evidence, while background regions and known objects can be more reliably distinguished.

We evaluate EviAttr-OW on the Real-World Object Detection (RWD) benchmark, which consists of datasets from five practical application scenarios. The experimental results show that our method achieves significant performance improvements and outperforms existing SOTA methods on all datasets. Notably, EviAttr-OW even surpasses the upper-bound open-vocabulary detection baseline that uses ground-truth category names, demonstrating stronger open-world generalization in terms of average precision on unknown categories. Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to reveal the impact of known-class semantic pull on unknown object decisions from the perspective of evidential attribute reasoning, and reformulate unknown object discovery as an interpretable evidential decision process.
- We propose EviAttr-OW, an evidential attribute reasoning framework for Open World Object Detection, and construct a class-agnostic attribute space that decouples attribute representations from known-class bias, thereby better covering potential unknown objects in the open world.
- We formulate unknown object discovery through the coupling of attribute responses and class-level evidence, deriving object evidence, evidential uncertainty, and Known-Class Evidence Deficiency for reliable unknown assignment.

- Experimental results on the RWD benchmark show that EviAttr-OW outperforms existing state-of-the-art methods on all datasets, achieving an absolute gain of 7.3 mAP in unknown object detection, which demonstrates stronger open-world generalization capability.

2 Related Works

2.1 Open World Object Detection

Open World Object Detection (OWOD) [6] extends object detection beyond the closed-set assumption, requiring detectors to discover unknown objects and incrementally learn new categories while recognizing known ones [7–10]. Existing methods mine unknown candidates via pseudo-label generation [11–13], objectness modeling [14, 15], hierarchical or contrastive representation learning [16, 17], and uncertainty-aware mining [18, 19]. These methods improve unknown recall, but often rely on known-class supervision or heuristic candidate scoring. Recent multimodal foundation models have motivated the adaptation of open-vocabulary object detection (OVD) [20–23] to OWOD through language-level semantic priors. FOMO [24] and PASS [25] further exploit large language models to derive attribute descriptions or semantic cues from known categories for unknown object discovery. However, since these cues are still generated from known-category semantics, the resulting attribute space remains category-dependent and may limit generalization to semantically diverse unknown objects. Different from these methods, EviAttr-OW reformulates unknown object discovery as an interpretable evidential decision process by converting class-agnostic attribute responses into Dirichlet evidence and combining object evidence with Known-Class Evidence Deficiency.

2.2 Evidential Deep Learning

Evidential Deep Learning (EDL) [26] formulates learning as a process of evidence acquisition, and introduces Dempster-Shafer theory [27] into uncertainty modeling for deep models. Unlike Bayesian or ensemble methods [28, 29], which rely on sampling or multiple model instances, EDL predicts the parameters of a Dirichlet distribution to deterministically characterize both categorical confidence and second-order uncertainty. Specifically, the model output is no longer interpreted as a point estimate of class probability, but as evidence supporting different class hypotheses. When sufficient evidence is accumulated, the model produces more reliable predictions; when evidence is scarce, the Dirichlet distribution naturally reflects higher uncertainty. Owing to its ability to jointly support uncertainty estimation and confidence calibration, EDL has been widely applied to open-set recognition [30, 31], multi-view classification [32–35], object detection [36, 37]. Inspired by the evidence acquisition principle in EDL, EviAttr-OW organizes the attribute responses of candidate regions into class-level evidence distributions to characterize known-class support and predictive uncertainty. Based on this evidential representation, we further construct Known-Class Evidence Deficiency, providing an interpretable evidence basis for unknown object decisions in open-world scenarios.

3 Methodology

As shown in Fig. 2, EviAttr-OW reconstructs attribute responses in Open World Object Detection from conventional classification signals into interpretable evidential representations. We define an unknown object as a candidate region that exhibits explicit object evidence while not being sufficiently explained by any known category. Based on this definition, EviAttr-OW proceeds in three steps: constructing a domain-relevant class-agnostic attribute space (Sec. 3.1), mapping attribute responses into Dirichlet evidence distributions via Evidential Attribute Modeling (EAM) (Sec. 3.2), and estimating unknown probabilities by combining object evidence with Known-Class Evidence Deficiency (Sec. 3.3).

3.1 Class-Agnostic Attribute Space Construction

OWOD requires discovering objects beyond the annotated category space. Although these objects have no category names, they often remain describable through visual attributes. This motivates a different construction principle: the attribute space should not be derived from known-category semantics, which tend to encode biased visual prototypes of known classes. Following this principle, EviAttr-OW constructs a class-agnostic attribute space from the task domain. Rather than asking

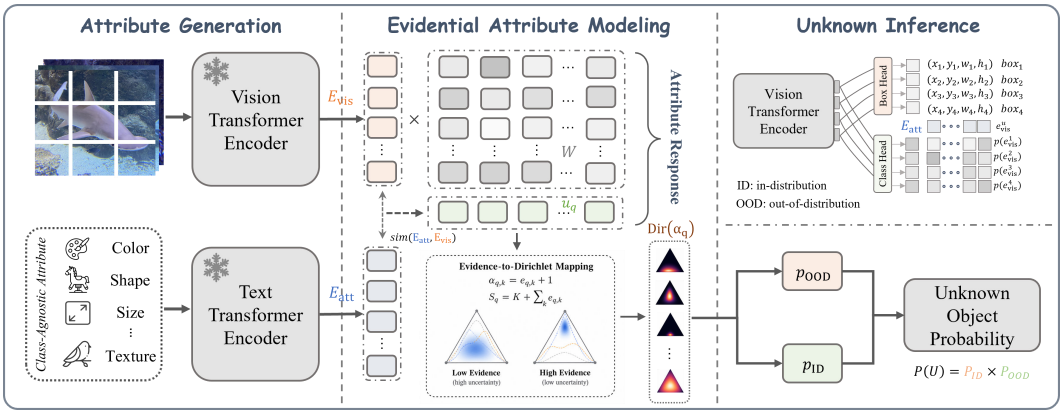

Figure 2: **Overall architecture of the proposed EviAttr-OW.** EviAttr-OW first constructs class-agnostic prompts and leverages LLM to generate domain-relevant visual attributes, which are encoded into E_{att} by a text encoder. Visual features E_{vis} are extracted from the input image and matched with these attributes to obtain fine-grained attribute responses. Evidential Attribute Modeling maps the attribute responses into class-level evidence and Dirichlet evidence distributions, producing known-class predictions, total evidence strength, and evidential uncertainty. Evidence-Calibrated Unknown Inference then combines object evidence p_{ID} with Known-Class Evidence Deficiency p_{OOD} to estimate the unknown object probability.

what attributes a specific known category has, we query general object attributes for the current domain, so that the resulting attribute space serves as a shared semantic basis for open-world object discovery instead of a category-dependent description of known classes.

Specifically, we adopt a set of general attribute types $\mathcal{Z}$, including shape, color, size, context, local part, overall appearance, environment, material, behavior, and texture. For each attribute type $z \in \mathcal{Z}$, we use a class-agnostic prompt template to query a large language model for attribute generation: "I am using a vision-language model to detect general objects. List typical z attributes, which will be used for detection." This yields a candidate attribute set $\mathcal{B}_z = \{b_{z,1}, b_{z,2}, \ldots, b_{z,n_z}\}$, where $b_{z,r}$ denotes the r-th attribute word generated for type z, and n_z is the number of attributes under this type. Each attribute pair $(z, b_{z,r})$ is then converted into a natural-language attribute description:

$$\text{Desc}(z, b_{z,r}) = \text{"object which is/has } z \text{ is } b_{z,r}\text{."} \tag{1}$$

All attribute descriptions are fed into a frozen text encoder $\phi_t(\cdot)$ to obtain attribute embeddings $\mathbf{e}^i_{att} \in \mathbb{R}^D$, which are stacked as

$$\mathbf{E}_{att} = [\mathbf{e}^1_{att}, \mathbf{e}^2_{att}, \ldots, \mathbf{e}^N_{att}]^\top \in \mathbb{R}^{N \times D}. \tag{2}$$

Given an input image, a frozen visual encoder $\phi_v(\cdot)$ extracts candidate-region or visual-token features. The visual embedding of the q-th candidate region is denoted as $\mathbf{e}^q_{vis} = \frac{\phi_v(q)}{\|\phi_v(q)\|_2} \in \mathbb{R}^D$. If the image contains Q candidate regions, their visual embeddings are stacked as

$$\mathbf{E}_{vis} = [\mathbf{e}^1_{vis}, \mathbf{e}^2_{vis}, \ldots, \mathbf{e}^Q_{vis}]^\top \in \mathbb{R}^{Q \times D}. \tag{3}$$

Thus, class-agnostic attributes and candidate regions are encoded into the same vision–text embedding space, enabling fine-grained semantic responses through attribute similarity.

3.2 Evidential Attribute Modeling

The class-agnostic attribute space provides a shared semantic basis for open-world objects, yet attribute activations alone do not yield reliable category decisions. Unknown objects may activate attributes shared by multiple known classes. Directly compressing such responses into closed-set probabilities forces the model to select a known category, leading to overconfident absorption of unknown objects. EAM therefore treats attribute responses as evidence rather than ordinary classification scores, and organizes them into a Dirichlet evidence distribution that captures class support, total evidence strength, and predictive uncertainty.

Given the visual embedding $\mathbf{e}_{vis}^{q} \in \mathbb{R}^{D}$ of candidate region q and the j-th attribute embedding $\mathbf{e}_{att}^{j} \in \mathbb{R}^{D}$, their normalized similarity is

$$\mathrm{sim}\left(\mathbf{e}_{vis}^{q}, \mathbf{e}_{att}^{j}\right) = \frac{\left(\mathbf{e}_{vis}^{q}\right)^{\top} \mathbf{e}_{att}^{j}}{\| \mathbf{e}_{vis}^{q} \|_2 \| \mathbf{e}_{att}^{j} \|_2}. \tag{4}$$

Since both embeddings are L_2-normalized, this similarity can be implemented as an inner product. The response of region q to the j-th attribute is defined as

$$u_{q,j} = \mathrm{Sigmoid}\left(\tau_a \cdot \mathrm{sim}\left(\mathbf{e}_{vis}^{q}, \mathbf{e}_{att}^{j}\right)\right), \tag{5}$$

where $\tau_a > 0$ is a temperature coefficient. To convert attribute responses into class-level evidence, we introduce a learnable attribute-to-class evidence mapping matrix $\mathbf{W} \in \mathbb{R}^{N \times K}$, where N is the number of attributes, K is the number of known categories, and $W_{j,k}$ denotes the evidence contribution of attribute a_j to category k. The aggregated evidence of candidate region q for category k is

$$e_{q,k} = \mathrm{Softplus}\left(\sum_{j=1}^{N} W_{j,k} u_{q,j}\right), \quad k = 1, 2, \ldots, K, \tag{6}$$

where $e_{q,k} \geq 0$ denotes the non-negative evidence supporting known category k. All class-level evidence values form an evidence vector:

$$\mathbf{e}_q = [e_{q,1}, e_{q,2}, \ldots, e_{q,K}] \in \mathbb{R}^{K}. \tag{7}$$

Unlike Softmax probabilities, $\mathbf{e}_q$ does not impose strict competition among categories, and therefore preserves the evidence strength of each known-class explanation. We further map class evidence to Dirichlet parameters:

$$\alpha_{q,k} = e_{q,k} + 1. \tag{8}$$

The constant $+1$ serves as a non-informative evidence prior. The total evidence strength is then defined as

$$S_q = \sum_{k=1}^{K} \alpha_{q,k} = K + \sum_{k=1}^{K} e_{q,k}. \tag{9}$$

When S_q is close to K, the region has limited known-class evidence; a larger S_q indicates stronger accumulated evidence. We obtain the known-class prediction and evidential uncertainty as

$$p_{q,k} = \frac{\alpha_{q,k}}{S_q}, \qquad U_q = \frac{K}{S_q}. \tag{10}$$

A larger U_q indicates weaker evidence and less reliable known-class explanation for subsequent unknown inference. To align the evidence distribution with known-category supervision, we use the evidence alignment loss:

$$\mathcal{L}_{align} = -\sum_{q} \sum_{k=1}^{K} y_{q,k} \log p_{q,k} = \sum_{q} \sum_{k=1}^{K} y_{q,k} \left(\log S_q - \log \alpha_{q,k}\right), \tag{11}$$

where $y_{q,k} \in \{0, 1\}$ indicates whether region q belongs to known category k. This loss increases target-category evidence while preventing unconstrained evidence growth through S_q. Open-world images contain background regions, ambiguous candidates, and potential unknown objects, which should not be overconfidently absorbed into known categories. We therefore introduce an isotropic evidence regularization term:

$$\mathcal{L}_{iso} = \sum_{q} \mathrm{KL}\left[\mathrm{Dir}\left(\widetilde{\boldsymbol{\alpha}}_q\right) \parallel \mathrm{Dir}\left(\mathbf{1}\right)\right], \tag{12}$$

where $\mathrm{Dir}(\mathbf{1})$ denotes the uniform Dirichlet prior, and $\widetilde{\boldsymbol{\alpha}}_q$ is the target-masked evidence parameter. For known-class samples, we define

$$\widetilde{\alpha}_{q,k} = y_{q,k} + (1 - y_{q,k})\alpha_{q,k}. \tag{13}$$

The masked parameter $\widetilde{\alpha}_{q,k}$ sets the target category to the non-informative prior while preserving non-target evidence, which is then regularized toward the uniform Dirichlet prior. This suppresses

excessive non-target evidence and prevents background or ambiguous regions from being absorbed into known categories. The training objective of EAM is

$$\mathcal{L}_{EAM} = \mathcal{L}_{align} + \lambda_{iso}\mathcal{L}_{iso}, \tag{14}$$

where λ_{iso} controls the strength of isotropic evidence regularization. The resulting Dirichlet distribution provides known-class predictions $p_{q,k}$, while S_q and U_q measure the sufficiency and reliability of the supporting evidence.

3.3 Evidence-Calibrated Unknown Inference

The class-level evidence distribution alone is insufficient for unknown object decisions, since an unknown object is neither simply the region with the lowest known-class probability nor the one with the highest attribute response. We consider a reliable unknown instance as a candidate region that exhibits clear object evidence while lacking sufficient known-class evidence. Accordingly, ECUI estimates the unknown probability by coupling object evidence with Known-Class Evidence Deficiency.

Given a candidate region q, we first use its strongest response in the attribute space to measure whether it activates valid object attributes, and define the attribute in-distribution probability as

$$p_q^{id} = \max_{j\in\{1,\dots,N\}} u_{q,j}, \tag{15}$$

where N denotes the number of foundational attributes. A larger p_q^{id} indicates that the region matches at least one foundational attribute and thus possesses potential object semantics.

However, the maximum response may be affected by accidental activations from background regions. We therefore introduce an empirical foreground probability to evaluate whether attribute activations form a stable object pattern. For an attribute response $u \in (0, 1)$, we define a shared calibration function and aggregate calibrated responses as

$$f(u) = \text{Sigmoid}\left(\beta_1 \log(u+\epsilon) + \beta_2 \log(1-u+\epsilon) + \beta_3\right), \tag{16}$$

$$p_q^{emp} = \frac{1}{N}\sum_{j=1}^{N} f(u_{q,j}), \tag{17}$$

where β_1, β_2, and β_3 are learnable calibration parameters, and $\epsilon > 0$ is a numerical stability constant. Compared with p_q^{id}, which depends only on the maximum response, p_q^{emp} captures whether the overall attribute activation is consistent with a foreground object, reducing false unknown predictions caused by isolated background responses.

Based on these components, the unknown probability of candidate region q is defined as

$$p_q^{unk} = \underbrace{\sqrt{p_q^{id} p_q^{emp}}}_{\substack{\text{Object Evidence}\\ P_{\text{ID}}(q)}} \cdot \underbrace{\left[1 - (1-U_q)\max_k p_{q,k}\right]}_{\substack{\text{Known-Class Evidence Deficiency}\\ P_{\text{OOD}}(q)}}. \tag{18}$$

where $P_{\text{ID}}(q)$ measures stable object evidence and $P_{\text{OOD}}(q)$ measures Known-Class Evidence Deficiency. The former suppresses accidental attribute activations from background regions, while the latter calibrates the strongest known-class prediction with evidential uncertainty. Thus, known objects are suppressed by sufficient known-class evidence, background regions are suppressed by insufficient object evidence, and real unknown objects obtain high unknown probabilities by jointly satisfying strong object evidence and weak known-class explanation.

4 Experiments

We provide detailed descriptions of the datasets, evaluation metrics, and implementation, along with additional ablation studies, in the Appendix.

Table 1: **Comparison with SOTA methods on the RWD benchmark.** We evaluate all methods on five real-world application domains with two incremental tasks per dataset. We report mAP for unknown classes (U), known classes (K), previously known classes (PK), and currently known classes (CK). Overall is computed by averaging results across all datasets and tasks. †GT denotes an oracle open-vocabulary baseline with access to ground-truth unknown category information.

Dataset	Aquatic				Aerial				Game				Medical				Surgery				Overall			
Task IDs (→)	Task 1		Task 2		Task 1		Task 2		Task 1		Task 2		Task 1		Task 2		Task 1		Task 2		Task 1		Task 2	
	U	K	PK	CK	U	K	PK	CK	U	K	PK	CK	U	K	PK	CK	U	K	PK	CK	U	K	PK	CK
B/16 Backbone:																								
BASE+ZS+GT†	29.8	45.0	45.0	36.7	1.3	5.7	5.7	1.4	15.0	0.4	0.4	0.1	0.5	0.0	0.0	0.1	5.6	1.5	1.4	0.3	10.4	10.5	10.5	7.7
BASE+ZS	6.2	45.0	45.0	36.7	0.9	5.7	5.7	1.4	15.7	0.4	0.4	0.1	0.2	0.0	0.0	0.1	1.4	1.5	1.4	0.3	4.9	10.5	10.5	7.7
BASE+ZS+IN	26.5	45.1	45.1	36.7	1.9	5.7	5.7	1.4	2.4	0.3	0.3	0.0	0.6	0.0	0.0	0.1	1.7	1.4	1.0	0.3	6.6	10.5	10.4	7.7
BASE+ZS+LLM	24.7	45.1	22.3	66.3	1.4	5.7	5.7	1.4	15.1	0.4	0.4	0.1	0.6	0.0	0.0	0.1	8.9	1.5	1.3	0.3	10.2	10.5	10.5	7.7
BASE+FS	7.1	41.1	41.1	31.9	1.2	10.4	10.1	4.0	16.0	4.6	4.8	3.9	0.6	6.1	6.1	3.3	1.3	11.9	11.3	10.9	5.2	14.8	14.7	10.8
UMB [38]	13.3	43.8	43.0	39.7	1.5	18.8	19.0	5.9	15.2	4.1	4.7	4.3	2.3	5.4	3.5	11.8	10.1	13.9	14.3	11.1	8.5	17.2	16.9	14.6
PASS [25]	5.2	43.4	43.2	46.6	1.9	14.0	16.0	7.0	21.5	10.0	7.7	9.0	4.9	8.4	6.8	12.1	14.3	15.6	13.1	14.7	9.6	18.3	17.4	17.9
FOMO [24]	3.5	43.8	44.1	40.8	0.9	12.0	12.6	5.4	13.3	3.8	4.4	4.1	2.1	6.4	5.5	11.5	6.1	12.7	12.9	11.0	5.2	15.7	15.9	14.6
Ours: EviAttr-OW	**16.3**	47.3	45.9	47.1	**2.8**	15.5	19.4	9.1	**27.4**	14.1	9.2	12.3	**9.8**	9.1	8.0	14.1	**18.2**	19.1	19.4	20.5	**14.9**	21.0	20.4	20.6
L/14 Backbone:																								
BASE+ZS+GT†	34.8	36.0	36.0	42.3	1.0	7.9	7.2	0.8	12.4	0.9	0.8	0.3	2.4	0.2	0.2	0.3	2.4	0.2	2.6	1.3	10.6	9.0	9.4	9.0
BASE+ZS	0.7	35.9	36.0	42.3	9.1	8.2	7.2	0.8	6.8	0.9	0.8	0.3	0.0	0.2	0.2	0.3	3.6	2.9	2.6	1.3	4.1	9.6	9.4	9.0
BASE+ZS+IN	19.6	35.8	35.8	41.8	2.3	7.2	6.9	0.9	15.8	0.9	0.8	0.3	0.0	0.2	0.2	0.3	3.1	2.1	1.9	1.1	8.3	9.2	9.1	8.8
BASE+ZS+LLM	24.7	35.8	35.8	42.2	0.6	7.6	7.2	0.8	12.5	0.9	0.8	0.2	1.6	0.1	0.1	0.2	12.6	2.6	2.5	1.3	10.4	9.4	9.3	9.0
BASE+FS	2.4	43.6	42.9	42.8	9.7	23.7	21.9	13.0	8.2	10.4	10.2	13.4	1.1	23.2	21.7	24.2	3.6	26.0	25.0	7.4	5.0	25.4	24.3	20.2
UMB [38]	18.6	50.7	50.5	50.4	11.2	42.7	40.4	22.6	35.1	11.1	10.7	10.5	13.2	22.2	19.1	34.5	24.5	36.6	39.0	17.4	20.5	32.7	31.9	27.1
PASS [25]	21.7	53.9	56.6	58.3	8.4	34.2	36.1	20.2	36.0	24.3	23.7	26.3	13.1	34.3	30.0	32.0	16.6	45.6	47.9	43.3	19.1	38.5	38.9	36.0
FOMO [24]	18.2	50.1	48.1	47.1	6.0	25.3	23.7	16.0	30.4	10.7	9.9	11.2	9.4	21.8	19.9	34.6	12.0	29.0	28.9	8.5	15.2	27.4	26.1	23.5
Ours: EviAttr-OW	**26.5**	55.8	58.4	60.3	**18.9**	44.1	41.6	22.7	**45.7**	26.2	26.9	29.0	**20.5**	38.3	38.6	36.4	**27.6**	48.2	50.4	45.8	**27.8**	42.5	43.2	38.8

Table 2: **Ablation study of EviAttr-OW.** We evaluate each variant on the RWD benchmark datasets under Task 1, reporting U-mAP (U) for unknown classes and K-mAP (K) for known classes. The consistent performance drops after ablating Class-Agnostic Attribute Space (CAS), Evidential Attribute Modeling (EAM), and Evidence-Calibrated Unknown Inference (ECUI) demonstrate the contribution of each component. More ablation results are provided in the supplementary material.

Dataset	Aquatic		Aerial		Game		Medical		Surgery		Overall	
	U	K	U	K	U	K	U	K	U	K	U	K
EviAttr-OW/CAS	22.8	53.2	13.6	39.5	41.5	22.8	17.3	35.2	24.8	44.9	24.0	39.1
EviAttr-OW/EAM	18.9	42.1	9.1	28.3	32.7	12.4	12.6	25.7	14.2	42.5	17.5	30.2
EviAttr-OW/ECUI	22.4	48.6	13.8	39.2	35.0	19.3	14.5	33.0	20.8	43.4	21.3	36.7
Final: EviAttr-OW	**26.5**	55.8	**18.9**	44.1	**45.7**	26.2	**20.5**	38.3	**27.6**	48.2	**27.8**	42.5

4.1 Datasets

Traditional OWOD evaluation protocols are commonly built upon VOC [39] and COCO [40]. However, in the era of vision foundation models, zero-shot detectors have shown increasingly saturated performance on these benchmarks, as their categories are mostly common and well represented in large-scale pretraining data. For example, OWL-ViT [23] has already achieved an unknown-class recall of 79.0 under the corresponding setting, suggesting that conventional benchmarks may no longer fully reflect the real challenges of open-world scenarios. Following FOMO [24], we evaluate our method on the RWD benchmark to better reflect practical application scenarios.

4.2 Comparison With SOTA Methods

We benchmark EviAttr-OW against representative OWOD methods, including UMB [38], PASS [25], and FOMO [24], together with open-vocabulary detector [23] based baselines adapted to the OWOD setting. These baselines span different levels of semantic or visual prior for unknown discovery. As shown in Tab. 1, BASE+ZS uses generic object prompts, BASE+ZS+IN and BASE+ZS+LLM expand the candidate space with ImageNet categories or LLM-predicted categories, BASE+FS uses a few visual exemplars with generic object prompts [41], and BASE+ZS+GT† serves as an oracle upper bound with access to ground-truth unknown category information.

Across five real-world datasets, EviAttr-OW achieves the best overall performance. In unknown object detection, EviAttr-OW reaches 27.8 mAP, outperforming PASS and UMB by 8.7 and 7.3 mAP, respectively. Compared with FOMO, EviAttr-OW achieves an 82.9% relative improvement, demonstrating stronger open-world generalization. Moreover, EviAttr-OW not only improves unknown-category

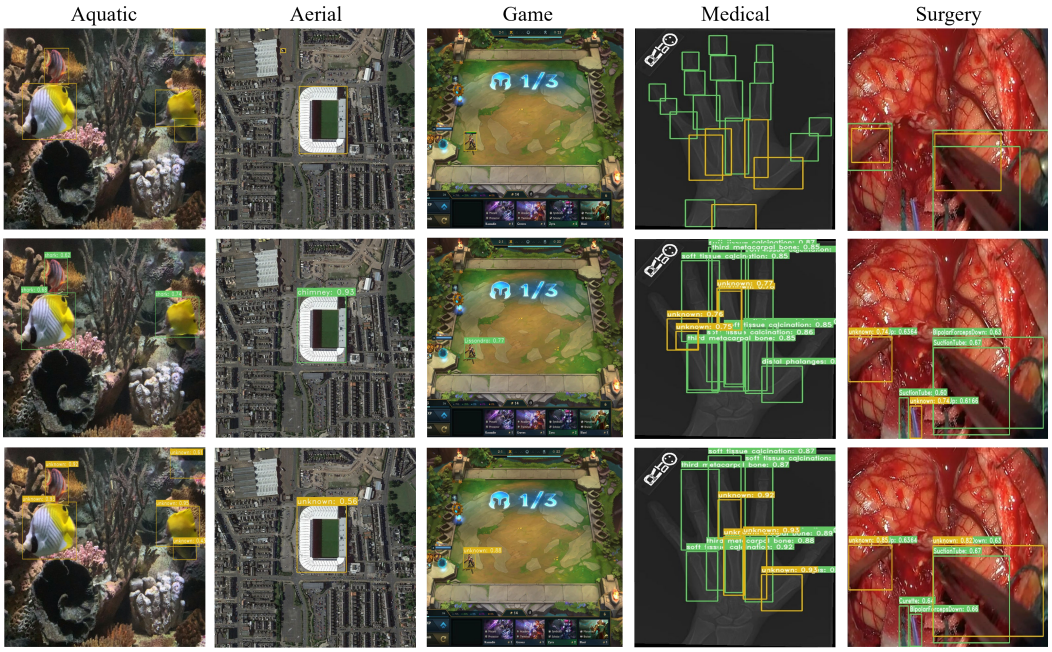

Figure 3: **Qualitative results and attribute-evidence interpretability.** Each column shows a representative example from one of the five RWD domains: Aquatic, Aerial, Game, Medical, and Surgery. The rows show (1) ground-truth annotations, (2) FOMO [24] detections, and (3) EviAttr-OW detections. Green and yellow boxes indicate known and unknown predictions, respectively.

detection but also enhances detection performance on known categories. For example, on known categories in Task 1, EviAttr-OW achieves 42.5 mAP, outperforming UMB by 9.8 points. These improvements stem from the reformulation of unknown decision logic in EviAttr-OW. EAM preserves the evidence structure within attribute responses, preventing potential unknown objects from being forcibly compressed into known-category probabilities. ECUI further combines object evidence with Known-Class Evidence Deficiency to select candidate regions that exhibit object attributes while lacking sufficient known-class support. This design improves the separability among unknown objects, known objects, and background regions, and provides clearer evidential explanations for unknown decisions. Although BASE+ZS+GT† uses the ground-truth category information of unknown classes and is theoretically regarded as the upper-bound reference for open-vocabulary detectors (OVD), its performance remains limited on several datasets. In contrast, EviAttr-OW does not rely on any prior knowledge of unknown categories, yet significantly outperforms BASE+ZS+GT† in multiple scenarios. For instance, BASE+ZS+GT† achieves only 2.4 mAP on both the Medical and Surgery datasets, whereas EviAttr-OW reaches 20.5 mAP and 27.6 mAP, respectively. These results indicate that simply relying on ground-truth category information does not guarantee open-world generalization. By contrast, EviAttr-OW can more effectively characterize potential unknown objects through evidential attribute reasoning.

4.3 Ablation Study

4.3.1 Effect of Class-Agnostic Attribute Space

The CAS module is designed to decouple attribute construction from known-category semantics, thereby reducing the semantic bias introduced by category-induced attributes. As shown in Tab. 2, removing CAS leads to consistent performance degradation across the RWD benchmark. This

result indicates that class-agnostic attribute construction plays an important role in improving the generalization of attribute representations to potential unknown objects. Without CAS, the attribute space tends to be more biased toward known categories, making unknown objects more likely to be absorbed into known-class semantics.

4.3.2 Effect of Evidential Attribute Modeling

The EAM module converts attribute responses into Dirichlet evidence distributions, allowing the model to explicitly represent class-level evidence, total evidence strength, and evidential uncertainty. As shown in Tab. 2, removing EAM causes a substantial performance drop among all variants. This demonstrates that directly using attribute responses without evidential modeling is insufficient for reliable open-world detection. By preserving the evidence structure of attribute responses, EAM prevents potential unknown objects from being forcibly compressed into closed-set category probabilities, thereby improving both unknown discovery and known-category recognition.

4.3.3 Effect of Evidence-Calibrated Unknown Inference

The ECUI module estimates unknown probability by combining object evidence with Known-Class Evidence Deficiency. This design selects candidate regions that exhibit clear object attributes while lacking sufficient support from known classes. As shown in Tab. 2, removing ECUI consistently degrades performance, confirming that unknown object discovery cannot be reliably achieved by attribute responses or class evidence alone. ECUI suppresses accidental attribute activations from background regions and prevents real unknown objects from being over-absorbed by known categories, leading to a more reliable evidential decision process for open-world detection.

4.4 Visualization

To evaluate the detection performance and attribute-evidence interpretability of EviAttr-OW, we compare it with FOMO [24] across five real-world domains in Fig. 3. FOMO exhibits missed detections, false detections, and inaccurate localization in challenging scenarios. In Aquatic, several unknown fish are missed or incompletely detected. In Aerial, a stadium-like unknown object is misclassified as a known category. In Game, the unknown character is detected with lower confidence and less accurate localization. In Medical, redundant predictions occur around similar anatomical structures. In Surgery, some surgical instruments and object regions are partially detected or missed. In contrast, EviAttr-OW achieves more reliable detection across these domains, improving unknown-object localization and recall while maintaining competitive known-category recognition. These results suggest that evidential attribute reasoning reduces confusion among known objects, unknown objects, and background regions. Moreover, EviAttr-OW provides interpretable attribute evidence for unknown predictions. As shown beneath each example in Fig. 3, the top attribute evidence captures semantic cues related to shape, appearance, context, material, and texture. Rather than merely providing confidence scores, these attributes reveal the visual evidence supporting unknown decisions. By grounding predictions in class-agnostic attribute evidence, EviAttr-OW enables more transparent and reliable unknown object discovery.

5 Conclusions

This paper addresses unknown object discovery in OWOD from an evidential attribute reasoning perspective. We propose EviAttr-OW, a novel framework that mitigates known-class semantic bias in attribute space and reformulates unknown detection as a joint decision based on object evidence and Known-Class Evidence Deficiency. Specifically, EAM preserves attribute evidence through Dirichlet modeling, while ECUI calibrates unknown probabilities by identifying regions with clear object evidence but insufficient known-class support. Experiments on the RWD benchmark demonstrate improved unknown detection and enhanced separability among unknown objects, known objects, and background regions. We hope this work contributes to more reliable and interpretable open-world object detection.

References

[1] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[2] Y. Sun, C. Zhang, X. Li, X. Jing, H. Kong, and Q.-G. Wang, "Mdsf-yolo: Advancing object detection with a multiscale dilated sequence fusion network," *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[3] Z. Huang, W. Li, X.-G. Xia, H. Wang, and R. Tao, "Task-wise sampling convolutions for arbitrary-oriented object detection in aerial images," *IEEE transactions on neural networks and learning systems*, vol. 36, no. 3, pp. 5204–5218, 2024.

[4] X. Jiang, T. Liu, L. Liu, Z. Gong, Y. Liu, and X. Li, "Policy generalization enhancement for uav active object detection via divide-and-conquer sharpness-aware gradient matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 36, no. 2, pp. 2273–2289, 2026.

[5] X. Hou, M. Liu, and S. Du, "Unfolddet: Advancing surface defect detection with dual feature separation and relation reasoning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 36, no. 4, pp. 5371–5383, 2026.

[6] K. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5830–5840.

[7] H. Huang, A. Geiger, and D. Zhang, "Good: Exploring geometric cues for detecting objects in an open world," *arXiv preprint arXiv:2212.11720*, 2022.

[8] Y. Li, Y. Wang, W. Wang, D. Lin, B. Li, and K.-H. Yap, "Open world object detection: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 2, pp. 988–1008, 2024.

[9] Y. Long, Y. Wen, J. Han, H. Xu, P. Ren, W. Zhang, S. Zhao, and X. Liang, "Capdet: Unifying dense captioning and open-world detection pretraining," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 233–15 243.

[10] M. Yang, J. Yin, Y. Gu, C. Deng, H. Zhang, and H. Zhu, "Consistent prompt tuning for generalized category discovery," *International Journal of Computer Vision*, pp. 1–28, 2025.

[11] A. Gupta, S. Narayan, K. Joseph, S. Khan, F. S. Khan, and M. Shah, "Ow-detr: Open-world detection transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9235–9244.

[12] Y. Ma, H. Li, Z. Zhang, J. Guo, S. Zhang, R. Gong, and X. Liu, "Annealing-based label-transfer learning for open world object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 454–11 463.

[13] Y. He, W. Chen, S. Wang, T. Liu, and M. Wang, "Recalling unknowns without losing precision: An effective solution to large model-guided open world object detection," *IEEE Transactions on Image Processing*, 2024.

[14] O. Zohar, K.-C. Wang, and S. Yeung, "Prob: Probabilistic objectness for open world object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 444–11 453.

[15] Z. Sun, J. Li, and Y. Mu, "Exploring orthogonality in open world object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 302–17 312.

[16] T. Doan, X. Li, S. Behpour, W. He, L. Gou, and L. Ren, "Hyp-ow: Exploiting hierarchical structure learning with hyperbolic distance enhances open world object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 1555–1563.

[17] S. Ma, Y. Wang, Y. Wei, J. Fan, T. H. Li, H. Liu, and F. Lv, "Cat: Localization and identification cascade detection transformer for open-world object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 681–19 690.

[18] H. Sarkar, V. Chudasama, N. Onoe, P. Wasnik, and V. N. Balasubramanian, "Open-set object detection by aligning known class representations," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 219–228.

[19] X. Xi, Y. Huang, J. Lin, and R. Luo, "Ktcn: Enhancing open-world object detection with knowledge transfer and class-awareness neutralization," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 2024, pp. 1462–1470.

[20] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li, "Learning to prompt for open-vocabulary object detection with vision-language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 084–14 093.

[21] C. Feng, Y. Zhong, Z. Jie, X. Chu, H. Ren, X. Wei, W. Xie, and L. Ma, "Promptdet: Towards open-vocabulary detection using uncurated images," in *European Conference on Computer Vision*. Springer, 2022, pp. 701–717.

[22] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 965–10 975.

[23] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen *et al.*, "Simple open-vocabulary object detection," in *European conference on computer vision*. Springer, 2022, pp. 728–755.

[24] O. Zohar, A. Lozano, S. Goel, S. Yeung, and K.-C. Wang, "Open world object detection in the era of foundation models," *arXiv preprint arXiv:2312.05745*, 2023.

[25] M. Yang, G. J. Goenawan, H. Qin, K. Han, X. Peng, Y. Yang, and H. Zhu, "Detecting open world objects via partial attribute assignment," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 318–20 328.

[26] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*, vol. 31, 2018.

[27] K. Sentz and S. Ferson, "Combination of evidence in dempster-shafer theory," 2002.

[28] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[29] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[30] W. Bao, Q. Yu, and Y. Kong, "Evidential deep learning for open set action recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 349–13 358.

[31] E. Aguilar, B. Raducanu, P. Radeva, and J. Van de Weijer, "Continual evidential deep learning for out-of-distribution detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3444–3454.

[32] W. Liu, X. Yue, Y. Chen, and T. Denoeux, "Trusted multi-view deep learning with opinion aggregation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 7, 2022, pp. 7585–7593.

[33] W. Liu, Y. Chen, and X. Yue, "Enhancing multi-view classification reliability with adaptive rejection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 18, 2025, pp. 18 969–18 977.

[34] ——, "Enhancing testing-time robustness for trusted multi-view classification in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 15 508–15 517.

[35] W. Liu, Y. Chen, X. Yue, C. Zhang, and S. Xie, "Enhancing reliability in medical image classification of imperfect views," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[36] R. Wang, R.-W. Zhao, X. Zhang, and R. Feng, "Towards evidential and class separable open set object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 6, 2024, pp. 5572–5580.

[37] Y. Yu, D. Deng, F. Liu, Q. Dou, Y. Jin, G. Chen, and P. A. Heng, "Anedl: Adaptive negative evidential deep learning for open-set semi-supervised learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 15, 2024, pp. 16 587–16 595.

[38] X. Xi, Y. Huang, Z. Zhong, and R. Luo, "Umb: Understanding model behavior for open-world object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 74 233–74 261, 2024.

[39] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[41] M. Maaz, H. Rasheed, S. Khan, F. S. Khan, R. M. Anwer, and M.-H. Yang, "Class-agnostic object detection with multi-modal transformer," in *European conference on computer vision*. Springer, 2022, pp. 512–531.

[42] F. Ciaglia, F. S. Zuppichini, P. Guerrie, M. McQuade, and J. Solawetz, "Roboflow 100: A rich, multi-domain object detection benchmark," *arXiv preprint arXiv:2211.13523*, 2022.

[43] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS journal of photogrammetry and remote sensing*, vol. 159, pp. 296–307, 2020.

[44] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele, and P. Jannin, "Detecting surgical tools by modelling local appearance and global shape," *IEEE transactions on medical imaging*, vol. 34, no. 12, pp. 2603–2617, 2015.

[45] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, "Objects365: A large-scale, high-quality dataset for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8430–8439.

[46] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

A Appendix / supplemental material

A.1 Datasets

We conduct experiments on the Real-World Object Detection (RWD) benchmark following FOMO [24]. Although the conventional OWOD benchmark built from COCO [40] and VOC [39] has been widely used in earlier studies, recent vision-language foundation models have shown increasingly strong zero-shot recognition ability on these datasets. As a result, this benchmark becomes less discriminative for evaluating unknown-object discovery in foundation-model-based OWOD settings. RWD provides a more challenging and practically oriented evaluation environment by covering five application domains with distinct visual characteristics. The Aquatic subset [42] contains underwater scenes where object appearance is often degraded by blur and illumination changes. The Aerial subset [43] includes remote-sensing images with small, dense, and visually similar objects. The Game subset [42] represents synthetic scenarios under limited data availability. The Medical subset [42] consists of X-ray images, where fine-grained category differences are difficult to distinguish without domain expertise. The Surgery subset [44] contains surgical tool images with severe occlusion and complex visual interference caused by blood and operating environments. Detailed category splits are reported in Tab. 3. Following the standard RWD protocol, each dataset is divided into two incremental tasks with disjoint category groups. During Task 1 evaluation, categories assigned to Task 2 are regarded as unknown, while the setting is reversed for Task 2. This split enables evaluation under a realistic open-world scenario, where the detector must recognize currently known categories while discovering categories that will be introduced in later stages.

A.2 Evaluation Metrics

We use mean Average Precision (mAP) as the primary evaluation metric for both known and unknown categories. For known categories, mAP is the standard metric in object detection and reflects the overall localization and classification quality. For unknown categories, earlier OWOD methods [6, 11] often report unknown recall to measure whether unseen objects can be discovered. However, recall alone does not penalize excessive false positives, and may favor detectors that assign many background regions to the unknown class. To provide a more balanced evaluation, we follow the recent RWD evaluation protocol and report unknown mAP for unknown-object detection. This metric considers both the recall of unknown objects and the precision of unknown predictions, making it more suitable for assessing open-world detection performance. Using mAP for both known and unknown categories also provides a unified evaluation criterion across all tasks and categories.

A.3 Implementation Details

All experiments are conducted on two NVIDIA RTX A800 GPUs. Following FOMO [24], we adopt frozen OWL-ViT [23] backbones, including L/14 and B/16, as the vision-language feature encoders. The detector is initialized from the model fine-tuned on a federated dataset composed of Objects365 [45] and Visual Genome [46], which provides strong cross-domain generalization for real-world object detection. For attribute construction, we use GPT-3.5 to generate class-agnostic visual attributes from domain-level prompts rather than known-category names. Specifically, for each attribute type, including shape, color, size, context, local features, appearance, environment, material, behavior, and texture, we query the language model to obtain domain-relevant attribute descriptions. These descriptions are then encoded by the frozen text encoder to construct the class-agnostic attribute embedding matrix E_{att}. Unless otherwise specified, we use $\tilde{N} = 25$ foundational attributes as the default setting. During training, visual embeddings E_{vis} are extracted from candidate regions using the frozen visual encoder and matched with the class-agnostic attribute embeddings to obtain attribute responses. The attribute-to-class evidence mapping matrix W and the calibration parameters in ECUI are optimized with known-category supervision. We train EviAttr-OW under the 100-shot setting of Task 1, where 100 labeled instances are available for each known class, while unknown classes remain unlabeled and inaccessible during training. For Evidential Attribute Modeling, we set the isotropic evidence regularization weight $\lambda_{iso} = 0.5$. Other hyperparameters, including the number of training epochs, optimizer, batch size, and learning rate, follow the configuration of FOMO [24] and are selected according to validation performance on known classes. The overall procedure of EviAttr-OW is summarized in Algorithm 1.

Algorithm 1 EviAttr-OW Training and Inference

Input: Image I with known labels $\{y_q\}$; attribute types $\mathcal{Z}$; frozen encoders ϕ_t, ϕ_v; learnable parameters $\mathbf{W}, \boldsymbol{\beta}$
Output: Unknown object probability p_q^{unk}
Function: $\texttt{CAS}(\mathcal{Z})$
for each attribute type $z \in \mathcal{Z}$ **do**
  $\mathcal{B}_z \leftarrow \mathrm{LLM}(\mathrm{Template}(z))$
  for each attribute word $b_{z,r} \in \mathcal{B}_z$ **do**
    $\mathrm{desc} \leftarrow \mathrm{Desc}(z, b_{z,r})$
    $\mathbf{e}_{att}^{j} \leftarrow \phi_t(\mathrm{desc})$
  end for
end for
return $\mathbf{E}_{att} = [\mathbf{e}_{att}^{1}, \mathbf{e}_{att}^{2}, \ldots, \mathbf{e}_{att}^{N}]^\top$
Function: $\texttt{EAM}(\mathbf{E}_{vis}, \mathbf{E}_{att}, \mathbf{W}, \{y_q\})$
for each candidate region q **do**
  for each attribute $j = 1, \ldots, N$ **do**
    $u_{q,j} \leftarrow \mathrm{Sigmoid}\left(\tau_a \cdot \mathrm{sim}(\mathbf{e}_{vis}^{q}, \mathbf{e}_{att}^{j})\right)$
  end for
  for each known category $k = 1, \ldots, K$ **do**
    $e_{q,k} \leftarrow \mathrm{Softplus}\left(\sum_{j=1}^{N} W_{j,k} u_{q,j}\right)$
    $\alpha_{q,k} \leftarrow e_{q,k} + 1$
  end for
  $S_q \leftarrow \sum_{k=1}^{K} \alpha_{q,k}, \quad p_{q,k} \leftarrow \alpha_{q,k}/S_q, \quad U_q \leftarrow K/S_q$
  $\mathcal{L}_{align}^{q} \leftarrow \sum_{k=1}^{K} y_{q,k} \left(\log S_q - \log \alpha_{q,k}\right)$
  $\widetilde{\alpha}_{q,k} \leftarrow y_{q,k} + (1 - y_{q,k})\alpha_{q,k}$
  $\mathcal{L}_{iso}^{q} \leftarrow \mathrm{KL}\left[\mathrm{Dir}(\widetilde{\boldsymbol{\alpha}}_q) \parallel \mathrm{Dir}(\mathbf{1})\right]$
end for
$\mathcal{L}_{EAM} \leftarrow \sum_q \mathcal{L}_{align}^{q} + \lambda_{iso} \sum_q \mathcal{L}_{iso}^{q}$
return $\{u_{q,j}\}, \{p_{q,k}\}, \{U_q\}, \mathcal{L}_{EAM}$
Function: $\texttt{ECUI}(\{u_{q,j}\}, \{p_{q,k}\}, U_q, \boldsymbol{\beta})$
for each candidate region q **do**
  $p_q^{id} \leftarrow \max_j u_{q,j}$
  for each attribute $j = 1, \ldots, N$ **do**
    $f(u_{q,j}) \leftarrow \mathrm{Sigmoid}\left(\beta_1 \log(u_{q,j} + \epsilon) + \beta_2 \log(1 - u_{q,j} + \epsilon) + \beta_3\right)$
  end for
  $p_q^{emp} \leftarrow \frac{1}{N} \sum_{j=1}^{N} f(u_{q,j})$
  $P_{\mathrm{ID}}(q) \leftarrow \sqrt{p_q^{id} p_q^{emp}}$
  $P_{\mathrm{OOD}}(q) \leftarrow 1 - (1 - U_q) \max_k p_{q,k}$
  $p_q^{unk} \leftarrow P_{\mathrm{ID}}(q) \cdot P_{\mathrm{OOD}}(q)$
end for
return $\{p_q^{unk}\}$

A.4 Detailed Attributes

Tab. 4 lists the ten visual attribute types adopted in EviAttr-OW. These types cover common object-related properties such as shape, color, size, context, appearance, material, behavior, and texture, forming a general attribute space for open-world perception. To ensure a fair comparison, we follow FOMO [24] in using the same high-level attribute types, but adopt a different attribute construction strategy. Instead of generating attributes from known class names, EviAttr-OW uses class-agnostic prompts that do not require category labels. This avoids making the attribute space overly dependent on known-class semantics and provides a more general basis for unknown-object discovery.

Table 3: **Dataset statistics.** Unknown (U) and known (K) categories in Task 1 correspond to currently known (CK) and previously known (PK) categories in Task 2, respectively. Numbers in parentheses denote the categories included in the OWL-ViT tokenizer vocabulary, and "in Tokenizer (%)" reports the corresponding coverage ratio.

Dataset (→)	**Aquatic**	**Aerial**	**Game**	**Medical**	**Surgery**
Train Images	318	5000	788	93	912
Test Images	319	5000	787	89	917
Unknown Classes	3 (3)	10 (8)	29 (3)	6 (1)	7 (2)
Known Classes	4 (4)	10 (8)	30 (5)	6 (0)	6 (0)
All Classes	7 (7)	20 (16)	59 (8)	12 (1)	13 (2)
in Tokenizer (%)	100	80.0	13.6	8.33	15.4

Table 4: **Attribute type descriptions.** Overview of the ten visual attribute types adopted in EviAttr-OW. Their textual descriptions define general object-related semantics and provide a class-agnostic attribute basis for open-world object detection.

Type	**Description**
Shape	This includes the overall form or outline of objects in the image
Color	The predominant colors present in the image and how they contribute to the class identity
Size	The dimensions and proportions of objects in the image
Context	The surroundings or background of objects, which can provide contextual information
Features	Specific visual characteristics unique to certain classes, like wings for birds or fins for fish
Appearance	The overall visual appearance including distinctive structures like wheels on vehicles or petals on flowers
Environment	The type of setting or habitat depicted in the image
Material	The substances or materials objects are made of, which can affect their appearance
Behavior	Attributes related to how objects in the image interact or move
Texture	The surface qualities of objects, such as smoothness, roughness, or patterns

A.5 Derivations and Properties of EviAttr-OW

A.5.1 Derivation of the Isotropic Dirichlet Regularizer

Open-world scenarios contain numerous background regions, ambiguous candidate regions, and potential unknown objects. These regions should not be over-confidently explained as any known class. Therefore, the main paper introduces an isotropic Dirichlet regularizer:

$$\mathcal{L}_{\text{iso}} = \sum_{q} \mathrm{KL}\left[\mathrm{Dir}(\widetilde{\boldsymbol{\alpha}}_q) \parallel \mathrm{Dir}(\mathbf{1})\right]. \tag{19}$$

where $\mathrm{Dir}(\mathbf{1})$ denotes the uniform Dirichlet distribution, corresponding to an uninformative state without explicit preference for any known class. For a known-class sample, the masked Dirichlet parameter is defined as

$$\widetilde{\alpha}_{q,k} = y_{q,k} + (1 - y_{q,k})\alpha_{q,k}. \tag{20}$$

This masking operation has a clear role. When $y_{q,k} = 1$, the target-class parameter is set to 1 and is thus not penalized by the regularizer. When $y_{q,k} = 0$, the non-target class parameter remains $\alpha_{q,k}$, so excessive evidence on non-target classes is constrained by the regularizer.

Let

$$\widetilde{S}_q = \sum_{k=1}^{K} \widetilde{\alpha}_{q,k}. \tag{21}$$

According to the KL divergence between two Dirichlet distributions, by substituting $\boldsymbol{\beta} = \mathbf{1}$, we obtain

$$\begin{aligned}
&\mathrm{KL}\left[\mathrm{Dir}(\widetilde{\boldsymbol{\alpha}}_q) \parallel \mathrm{Dir}(\mathbf{1})\right] \\
&= \underbrace{\log\Gamma(\widetilde{S}_q) - \log\Gamma(K)}_{\text{global concentration mismatch}} - \underbrace{\sum_{k=1}^{K} \log\Gamma(\widetilde{\alpha}_{q,k})}_{\text{class-wise evidence normalization}} + \underbrace{\sum_{k=1}^{K} (\widetilde{\alpha}_{q,k} - 1)\left[\psi(\widetilde{\alpha}_{q,k}) - \psi(\widetilde{S}_q)\right]}_{\text{non-target evidence deviation}}.
\end{aligned} \tag{22}$$

where $\Gamma(\cdot)$ denotes the Gamma function, and $\psi(\cdot)$ denotes the digamma function. The above formulation can be understood as three components. The first component measures the mismatch between the global concentration strength and the uniform prior. The second component characterizes the class-wise evidence normalization. The third component directly measures the deviation of class parameters from the uniform prior.

Since the ground-truth class parameter is masked to 1, this KL regularizer mainly acts on non-target class evidence. In other words, it does not weaken the evidence learning of the ground-truth class, but suppresses excessive evidence accumulation on incorrect classes. This design is crucial for open-world detection, because background regions and unknown objects may activate certain attributes related to known classes. Without such regularization, the model may over-confidently explain these regions as known classes. The isotropic Dirichlet regularizer reduces the risk of incorrectly absorbing unknown objects and ambiguous candidate regions into known classes by pushing non-target evidence toward an uninformative prior.

A.5.2 Evidence-Uncertainty Decomposition

In the main paper, EviAttr-OW converts class evidence into Dirichlet parameters and further defines predictive uncertainty U_q through the total evidence strength S_q. To clarify why this uncertainty can characterize the degree to which a candidate region cannot be explained by the current known label space, we provide the following evidence-uncertainty decomposition.

For candidate region q, we define the belief mass of class k as

$$b_{q,k} = \frac{e_{q,k}}{S_q}. \tag{23}$$

Meanwhile, the predictive uncertainty is defined as

$$U_q = \frac{K}{S_q}. \tag{24}$$

Since $S_q = K + \sum_{k=1}^{K} e_{q,k}$, we have

$$\underbrace{\sum_{k=1}^{K} b_{q,k}}_{\text{explained known-class belief}} + \underbrace{U_q}_{\text{unexplained uncertainty}} = \sum_{k=1}^{K} \frac{e_{q,k}}{S_q} + \frac{K}{S_q} = \frac{\sum_{k=1}^{K} e_{q,k} + K}{S_q} = 1. \tag{25}$$

This relationship shows that the prediction state of a candidate region can be fully decomposed into two parts: the belief mass explained by known classes and the uncertainty unexplained by the current known label space. Unlike ordinary Softmax probabilities, the Dirichlet evidential distribution not only produces class predictions but also explicitly preserves how much evidence supports the prediction.

Furthermore, when a candidate region lacks effective class evidence, we have

$$e_{q,k} \approx 0,\ \forall k, \quad S_q \approx K, \quad U_q \approx 1. \tag{26}$$

This means that the model cannot form a reliable explanation for the candidate region using the current known classes. In contrast, when the candidate region accumulates a large amount of class evidence:

$$\sum_{k=1}^{K} e_{q,k} \to \infty, \quad S_q \to \infty, \quad U_q \to 0. \tag{27}$$

Therefore,

$$\lim_{\sum_k e_{q,k}\to 0} U_q = 1, \quad \lim_{\sum_k e_{q,k}\to \infty} U_q = 0. \tag{28}$$

This property shows that U_q is inversely related to the total evidence strength. Weaker evidence leads to stronger uncertainty, while sufficient evidence leads to lower uncertainty. This provides a theoretical basis for using U_q to calibrate known-class evidence in unknown inference.

This decomposition also explains why relying solely on the maximum class probability $\max_k p_{q,k}$ is insufficient for unknown prediction. Since the Dirichlet predictive mean is influenced by both class evidence and the uninformative prior, a candidate region may still obtain a relatively high class probability for some known class even when the supporting evidence is weak. Therefore, a high class probability does not necessarily indicate reliable class evidence. EviAttr-OW uses $(1-U_q)\max_k p_{q,k}$ to measure reliable known-class evidence, thereby filtering out unreliable predictions with high probability but weak evidence.

A.5.3 Derivation of the Empirical Foreground Calibration Function

In the main paper, EviAttr-OW introduces the empirical foreground probability p_q^{emp} to reduce the risk that background regions are misclassified as unknown objects due to accidental activation of a single attribute. This section explains the origin of this calibration function from a likelihood-ratio perspective.

Let $u \in (0,1)$ denote an attribute response. Since foreground objects and background regions usually exhibit different attribute-response distributions, we assume that the attribute response follows two Beta distributions under the foreground and background hypotheses:

$$u \mid \mathrm{fg} \sim \mathrm{Beta}(a_f, b_f), \quad u \mid \mathrm{bg} \sim \mathrm{Beta}(a_b, b_b). \tag{29}$$

According to Bayes' rule, the posterior probability of foreground is

$$P(\mathrm{fg} \mid u) = \frac{P(u \mid \mathrm{fg})P(\mathrm{fg})}{P(u \mid \mathrm{fg})P(\mathrm{fg}) + P(u \mid \mathrm{bg})P(\mathrm{bg})}. \tag{30}$$

Taking the logit transformation of this posterior probability gives

$$\log \frac{P(\mathrm{fg} \mid u)}{1 - P(\mathrm{fg} \mid u)} = \log \frac{P(u \mid \mathrm{fg})P(\mathrm{fg})}{P(u \mid \mathrm{bg})P(\mathrm{bg})}. \tag{31}$$

The density of the Beta distribution is

$$P(u \mid a, b) = \frac{1}{B(a,b)} u^{a-1}(1-u)^{b-1}. \tag{32}$$

Substituting the foreground and background Beta densities into the likelihood ratio yields

$$\begin{aligned} \log \frac{P(\mathrm{fg} \mid u)}{1 - P(\mathrm{fg} \mid u)} = & \underbrace{(a_f - a_b)\log u}_{\text{response activation contrast}} + \underbrace{(b_f - b_b)\log(1-u)}_{\text{response suppression contrast}} \\ & + \underbrace{\log \frac{P(\mathrm{fg})B(a_b, b_b)}{P(\mathrm{bg})B(a_f, b_f)}}_{\text{foreground-background prior contrast}}. \end{aligned} \tag{33}$$

Let

$$\beta_1 = a_f - a_b, \quad \beta_2 = b_f - b_b, \quad \beta_3 = \log \frac{P(\mathrm{fg})B(a_b, b_b)}{P(\mathrm{bg})B(a_f, b_f)}. \tag{34}$$

Then,

$$P(\mathrm{fg} \mid u) = \mathrm{Sigmoid}\left(\underbrace{\beta_1 \log u}_{\text{activation term}} + \underbrace{\beta_2 \log(1-u)}_{\text{suppression term}} + \underbrace{\beta_3}_{\text{bias term}} \right). \tag{35}$$

Considering numerical stability, the empirical foreground calibration function in the main paper can be written as

$$f(u) = \mathrm{Sigmoid}\left(\beta_1 \log(u+\epsilon) + \beta_2 \log(1-u+\epsilon) + \beta_3\right). \tag{36}$$

Therefore, $f(u)$ can be interpreted as a foreground posterior estimate under a Beta likelihood-ratio model, rather than an arbitrarily designed nonlinear mapping. The empirical foreground probability of candidate region q is obtained by averaging the calibrated responses over all attributes:

$$p_q^{emp} = \frac{1}{N} \sum_{j=1}^{N} f(u_{q,j}). \tag{37}$$

This design complements the maximum attribute response p_q^{id}. The term p_q^{id} focuses on whether the candidate region strongly activates at least one basic attribute, while p_q^{emp} focuses on whether the candidate region exhibits a stable foreground pattern in the overall attribute space. If a background region only accidentally activates a single attribute, its maximum attribute response may be high, but its overall empirical foreground probability is usually low. Therefore, combining these two terms reduces the risk of misclassifying background regions as unknown objects.

A.5.4 Properties of Evidence-Calibrated Unknown Inference

In the main paper, EviAttr-OW models the unknown probability as the product of Object Evidence and Known-Class Evidence Deficiency:

$$p_q^{unk} = \underbrace{P_{\mathrm{ID}}(q)}_{\text{Object Evidence}} \cdot \underbrace{P_{\mathrm{OOD}}(q)}_{\text{Known-Class Evidence Deficiency}}. \tag{38}$$

Specifically,

$$P_{\mathrm{ID}}(q) = \sqrt{p_q^{id} p_q^{emp}}, \quad P_{\mathrm{OOD}}(q) = 1 - (1 - U_q) \max_k p_{q,k}. \tag{39}$$

Let

$$p_q^{\max} = \max_k p_{q,k}. \tag{40}$$

Then,

$$P_{\mathrm{OOD}}(q) = 1 - \underbrace{(1 - U_q) p_q^{\max}}_{\text{Reliable known-class evidence}}. \tag{41}$$

where $p_q^{\max}$ denotes the strongest known-class prediction, and $1 - U_q$ denotes its evidential reliability. Therefore, $(1 - U_q) p_q^{\max}$ measures the degree to which the candidate region can be reliably explained by known classes. Accordingly, $P_{\mathrm{OOD}}(q)$ measures its Known-Class Evidence Deficiency.

Monotonicity analysis. Taking the derivative of $P_{\mathrm{OOD}}(q)$ with respect to U_q gives

$$\frac{\partial P_{\mathrm{OOD}}(q)}{\partial U_q} = p_q^{\max} \geq 0. \tag{42}$$

This shows that, when the maximum known-class prediction is fixed, higher evidential uncertainty increases Known-Class Evidence Deficiency. In other words, if the model lacks sufficient evidence to explain a candidate region with known categories, its unknown tendency should increase.

Taking the derivative of $P_{\mathrm{OOD}}(q)$ with respect to $p_q^{\max}$ gives

$$\frac{\partial P_{\mathrm{OOD}}(q)}{\partial p_q^{\max}} = -(1 - U_q) \leq 0. \tag{43}$$

This shows that, when uncertainty is fixed, a stronger maximum known-class prediction reduces Known-Class Evidence Deficiency. In particular, when U_q is low, the model has reliable known-class evidence, and a high $p_q^{\max}$ strongly suppresses the unknown probability. When U_q is high, the prediction lacks evidential support, and even a relatively high $p_q^{\max}$ has a weakened suppressive effect.

This property highlights the key difference between EviAttr-OW and maximum-probability-based decision rules. Ordinary methods only focus on $\max_k p_{q,k}$, while EviAttr-OW further evaluates whether this maximum probability is supported by sufficient evidence.

Boundedness analysis. Since attribute responses are obtained by the Sigmoid function, and the empirical foreground calibration function is also a Sigmoid output, we have

$$u_{q,j} \in (0,1), \quad p_q^{id} \in (0,1), \quad f(u_{q,j}) \in (0,1), \quad p_q^{emp} \in (0,1). \tag{44}$$

Therefore,

$$P_{\mathrm{ID}}(q) = \sqrt{p_q^{id} p_q^{emp}} \in (0,1). \tag{45}$$

On the other hand, since $U_q \in [0,1]$ and $p_q^{\max} \in [0,1]$, we have

$$(1-U_q)p_q^{\max} \in [0,1], \quad P_{\mathrm{OOD}}(q) = 1-(1-U_q)p_q^{\max} \in [0,1]. \tag{46}$$

Finally,

$$p_q^{unk} = P_{\mathrm{ID}}(q) P_{\mathrm{OOD}}(q) \in [0,1]. \tag{47}$$

This property shows that the unknown score of EviAttr-OW has a stable probabilistic range and will not become numerically uncontrolled due to variations in evidence strength or attribute responses. Meanwhile, the product form requires the unknown probability to satisfy two conditions simultaneously: the candidate region should have object evidence and exhibit Known-Class Evidence Deficiency. If either condition is not satisfied, the unknown probability will be suppressed.

A.6 Additional Experiments

A.6.1 Effect of Foundational Attribute Number

We further analyze the effect of the number of foundational attributes $\tilde{N}$ used in EviAttr-OW. Since the class-agnostic attribute space provides the semantic basis for evidential attribute reasoning, an appropriate attribute budget is important for balancing semantic coverage and noise suppression. As shown in Tab. 5, using a small number of attributes, such as $\tilde{N} = 5$ or $\tilde{N} = 10$, already yields competitive unknown-category performance, indicating that EviAttr-OW can extract useful evidence from a compact attribute set. However, limited attributes may not sufficiently cover diverse visual semantics, leading to lower known-category performance. When $\tilde{N} = 25$, EviAttr-OW achieves the best overall results on both Aquatic and Surgery, especially improving known-category detection while maintaining strong unknown-category performance. This suggests that a moderately sized attribute space provides richer and more stable evidence for both known-class recognition and unknown object discovery. Increasing the attribute number to $\tilde{N} = 50$ does not further improve performance and even leads to degradation, likely because excessive attributes introduce redundant or noisy responses into the evidence aggregation process. Therefore, we adopt $\tilde{N} = 25$ as the default setting, which offers a favorable trade-off between attribute diversity and evidential reliability.

A.6.2 Different Choices of P_{ID}

To further analyze the design of object evidence in ECUI, we compare different choices of P_{ID} while fixing P_{OOD} as our Known-Class Evidence Deficiency. As shown in Tab. 6, conventional aggregation strategies, including Max, Mean, Mahalanobis distance (Mah.), and Max+Mean, achieve similar performance on known categories but provide limited gains for unknown detection. This indicates that simply relying on the strongest attribute response, averaging all responses, or measuring distributional distance is insufficient to robustly distinguish unknown objects from known categories and background regions. In contrast, our proposed object evidence achieves the best unknown detection performance, improving unknown mAP to 27.6 while maintaining stable known-category performance. Compared with the best conventional alternative, Mah., our design brings a clear improvement of 5.6 mAP on unknown categories. Moreover, compared with the variant without empirical foreground estimation, the full design further improves unknown mAP from 25.1 to 27.6, confirming the importance of modeling stable foreground-like attribute activation. These results demonstrate that reliable unknown inference requires not only a strong attribute response, but also an empirical estimation of whether the response pattern corresponds to a real foreground object. By combining the maximum attribute response with empirical foreground evidence, our P_{ID} provides a more robust object-evidence term for ECUI.

Table 5: **Effect of Foundational Attribute Number.** We present U-mAP (U) and K-mAP (K) results on two representative datasets, Aquatic and Surgery. All experiments follow the Task 1 setting with 100 shots per class. The variable $\tilde{N}$ indicates the number of foundational attributes used during training.

Dataset (→)	**Aquatic**		**Surgery**	
	U	**K**	**U**	**K**
EviAttr-OW ($\tilde{N}$ = 5)	25.4	51.2	24.9	39.9
EviAttr-OW ($\tilde{N}$ = 10)	25.7	53.3	25.6	42.8
EviAttr-OW ($\tilde{N}$ = 25, default)	**26.5**	**55.8**	**27.6**	**48.2**
EviAttr-OW ($\tilde{N}$ = 50)	24.8	43.2	26.0	47.5

Table 6: **Experimental results of ECUI with different P_{ID} choices.** The evaluation is conducted under the same setting while fixing P_{OOD} as our Known-Class Evidence Deficiency on Surgery. The results show that our object evidence yields better unknown detection while maintaining competitive known-class performance.

p_{ID} (→)	**Max**	**Mean**	**Mah**	**Max+Mean**	w/o **emp**	**Ours**
known-mAP	48.2	48.1	48.3	48.2	48.2	48.2
UnKnown-mAP	21.8	21.6	22.0	21.8	25.1	27.6

A.6.3 Failure Cases

Fig. 4 presents representative failure cases from the Aquatic, Aerial, Game, Medical, and Surgery datasets, focusing on the challenges of unknown object recall and false positive suppression. Overall, EviAttr-OW demonstrates superior robustness and semantic discrimination capability compared to FOMO under complex real-world scenarios. In the Aquatic dataset, FOMO fails to accurately capture the complete contours of two fish. One fish is entirely missed and incorrectly treated as background, while the other is detected with a noticeably shifted bounding box, which substantially reduces recall. In contrast, EviAttr-OW produces tighter localizations of the fish bodies, although it misses a known-category prediction and shows slight boundary drift. In the Aerial dataset, FOMO mistakenly classifies rooftop structures in the background as unknown objects, whereas EviAttr-OW successfully identifies the true unknown instance but introduces overlapping predictions.

A.7 Limitations

Our focus is on converting attribute responses into interpretable evidential representations for OWOD. Since EviAttr-OW does not directly redesign the underlying vision-language detector but mainly reasons over its candidate outputs, the performance ceiling of our method can still be constrained by the inherent capability of the base detector. Exploring more advanced base detectors and more efficient attribute-evidence reasoning strategies could further extend the applicability of EviAttr-OW.

A.8 Broader Impact

EviAttr-OW aims to improve open-world object detection by discovering unknown objects with interpretable attribute evidence. This may help reduce annotation cost, improve the transparency of model decisions, and benefit applications such as robotics, remote sensing, medical image analysis, and industrial inspection. As with general object detection techniques, potential misuse in privacy-sensitive scenarios such as surveillance should be considered. Therefore, practical deployment should involve domain-specific validation, human oversight, and appropriate privacy and safety safeguards.

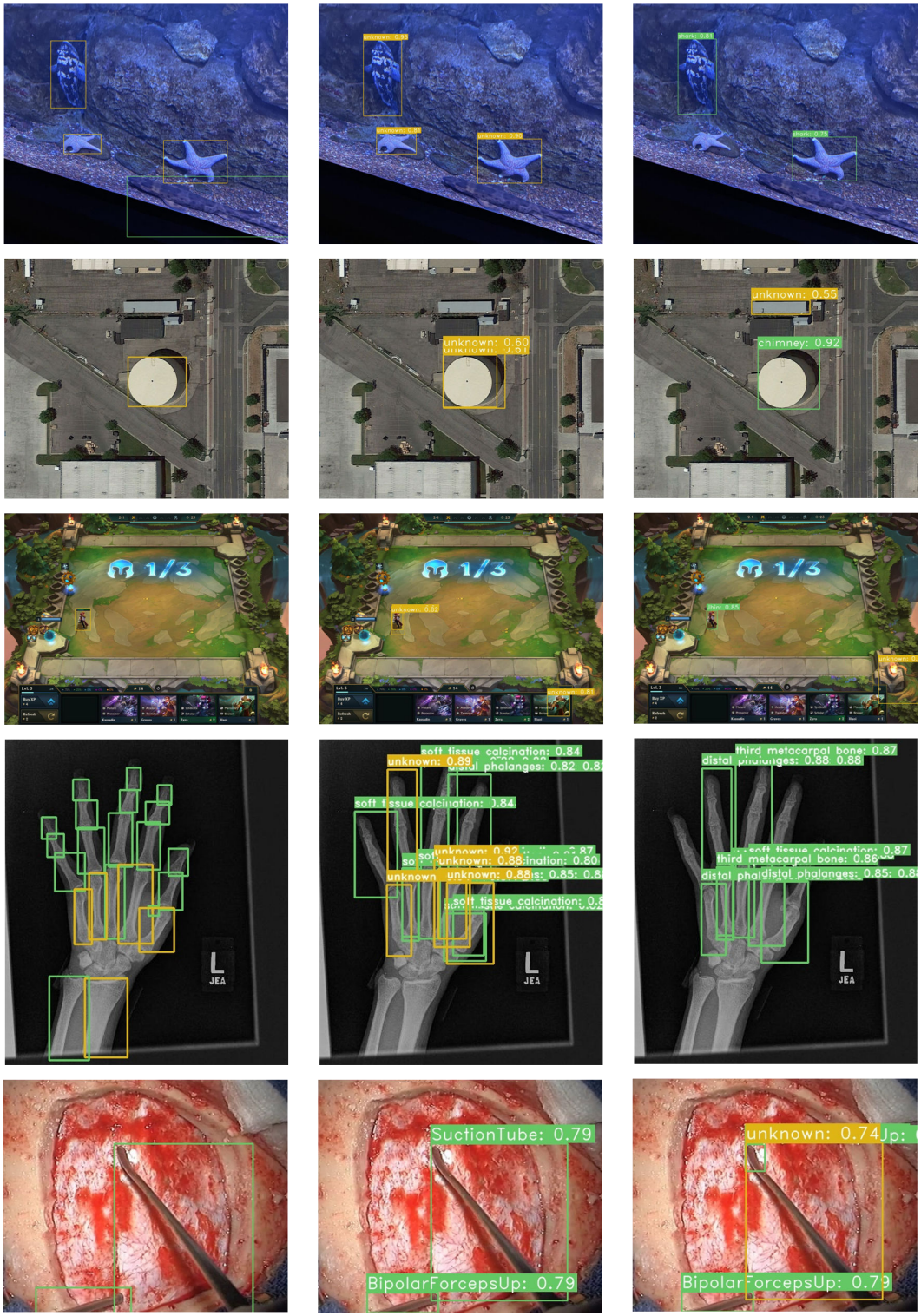

Figure 4: **Qualitative Results (Failure Cases).** From top to bottom, examples are selected from the Aquatic, Aerial, Game, Medical, and Surgery datasets. For clarity, we visualize representative high-confidence predictions for both known and unknown categories. Compared with FOMO, EviAttr-OW produces more reliable unknown-object predictions in these challenging cases, with fewer missed detections and improved localization accuracy.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The main claims in the abstract and introduction accurately reflect the contributions and scope of the paper. Specifically, the paper claims that EviAttr-OW discovers unknown objects through class-agnostic attribute construction, evidential attribute modeling, and evidence-calibrated unknown inference, which are described in Sec. 3 and supported by the experimental results in Sec. 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes].

Justification: We discuss the limitations of EviAttr-OW in Appendix A.7. In particular, we analyze its dependence on the quality of candidate regions, the coverage of the class-agnostic attribute space, the reliability of vision-language representations, and the computational cost introduced by attribute-based evidence reasoning.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: The paper provides mathematical formulations and property analyses for the proposed evidential attribute reasoning mechanism in Sec. 3 and Appendix A.5. The appendix further derives the isotropic Dirichlet regularizer, evidence-uncertainty decomposition, empirical foreground calibration, and properties of evidence-calibrated unknown inference.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: We provide the information needed to reproduce the main experimental results, including benchmark construction, task splits, evaluation metrics, model architecture, training objectives, implementation details, and hyperparameters in Sec. 4 and Appendix A.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: We do not provide the code at submission time because the repository is still being anonymized and cleaned. Nevertheless, we provide detailed descriptions of the method, data splits, evaluation metrics, hyperparameters, implementation details, and algorithmic pseudocode in the paper and supplemental material to support reproducibility. The code will be released upon publication.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The paper specifies the experimental setting, including the RWD benchmark, task splits, evaluation metrics, compared methods, backbone models, few-shot setting, attribute construction protocol, and implementation details in Sec. 4 and Appendix A.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No].

   Justification: Consistent with prior OWOD studies, our evaluation follows fixed benchmark splits and predefined incremental task settings, where statistical significance tests or error bars are not commonly reported. To strengthen the reliability of the empirical findings, we provide extensive comparisons against representative baselines, systematic component ablations, and qualitative analyses across five real-world domains.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: We provide the main compute resources used in our experiments in Appendix A.3, including the GPU type and number of GPUs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that the research conducted in this paper conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: We discuss the societal impacts of the proposed method in Sec. A.8, including both potential benefits and risks.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate

to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release high-risk assets such as pretrained language models, image generators, or newly scraped datasets. The released code and models, if any, are intended for research on public object detection benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: All existing assets used in this work, including datasets, pretrained models, and codebases, are properly cited, and their licenses and terms of use are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: We release the implementation of EviAttr-OW with documentation, configuration files, and instructions for reproducing the main experiments. The released assets are anonymized for submission and will be made publicly available upon acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing, user studies, or human-subject research, and therefore IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes].

    Justification: Following prior OWOD methods that leverage LLMs for attribute construction, we use GPT-3.5 to assist in generating visual attributes. These attributes are then encoded by a frozen text encoder and incorporated into the proposed evidential attribute reasoning pipeline.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.