MGME: MANIFOLD-GUIDED MULTI-PROTOTYPE EXPANSION FOR OPEN WORLD OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In Open World Object Detection, a detector must recognize known classes while simultaneously identifying instances of unknown categories. Existing approaches typically represent unknown objects using the arithmetic mean of known class embeddings or a single generic text prompt. However, a single mean proxy compresses diverse semantic directions into one representation and may fall into semantically ambiguous regions, a limitation we term *mean collapse*, thereby restricting its ability to capture multimodal unknown semantics. To overcome this geometric flaw, we propose the **Manifold-Guided Multi-Prototype Expansion (MGME)** framework. Instead of compressing unknowns into a singular proxy, MGME models them as a diverse set of unknown prototypes distributed along the valid boundaries of known classes via orthogonal geodesic traversal. Specifically, we extract a focused orthogonal residual subspace to serve as a collision-free exploration space for novelty generation. Furthermore, we design a dual-constraint selection mechanism that combines cross-class boundary pruning with large language model-driven semantic steering to ensure the generated prototypes are both topologically safe and semantically aligned. Finally, we introduce Discriminative Prototype Orthogonalization during the training phase to push known text prototypes apart, establishing highly discriminative decision boundaries. Extensive experiments across five real-world object detection datasets demonstrate that MGME effectively prevents manifold collisions and generates novel unknown prototypes with strong visual objectness, achieving state-of-the-art performance and yielding significant improvements in unknown object discovery.

1 INTRODUCTION

Open World Object Detection (OWOD) represents a critical step toward artificial general intelligence in vision, requiring models to simultaneously identify known classes and discover previously unseen objects without prior category information Joseph et al. (2021); Gupta et al. (2022); Zohar et al. (2023b); Zhao et al. (2023); Li et al. (2024b). Recently, the rise of multi-modal foundation models has catalyzed Open-Vocabulary Object Detection (OVOD). By leveraging rich vision-language alignment, OVOD methods such as ViLD Gu et al. (2021), RegionCLIP Zhong et al. (2022), and F-VLM Kuo et al. (2022) have achieved remarkable zero-shot generalization to novel concepts. However, these frameworks inherently rely on explicit text prompts to localize targets, making them unsuitable for the strict OWOD setting, where a system must discover true "unknowns" in the absence of any prior metadata regarding the unseen categories Joseph et al. (2021); Zohar et al. (2023a); Xi et al. (2024).

Existing attribute-driven OWOD frameworks Zohar et al. (2023a); Xi et al. (2024); Yang et al. (2025) attempt to bridge this gap by decomposing abstract categories into class-agnostic attributes, yet they encounter a critical bottleneck when mathematically modeling the representation of the unbounded unknown space. This limitation can be understood from a geometric perspective: previous methodologies predominantly operate under an implicit flat Euclidean assumption when constructing representations for the unknown category, as illustrated in Fig. 1(a), while the feature space of vision-language models is a hyperspherical Riemannian manifold with constant sectional curvature, the distribution of embeddings within this space is often highly anisotropic Radford et al. (2021); Ming et al. (2022). Performing naive Euclidean interpolation within such an anisotropic distribution can lead to *mean collapse*: the resulting unknown prototypes move toward low-density regions and

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



(a) Previous Method. (b) Our manifold-guided multi-prototype expansion.

Figure 1: Comparison of feature-space representations. Previous methods construct a single mean-based unknown representation, whereas our method generates multiple unknown prototypes through manifold-guided expansion.

lose the semantic diversity preserved by the hyperspherical manifold. Consequently, these mean-based proxies fall into semantically ambiguous regions, failing to capture the multimodal diversity of real-world out-of-distribution (OOD) targets Liang et al. (2017); Sun et al. (2021).

To address these fundamental deficiencies, we propose the **Manifold-Guided Multi-Prototype Expansion (MGME)** framework, which reformulates unknown modeling from a rigorous Riemannian perspective. Rather than aggregating known embeddings into a collapsed mean, MGME dynamically identifies an *orthogonal residual subspace* that acts as a semantic vacuum devoid of known class variance while preserving structural objectness. Within this subspace, we execute principled *geodesic traversals* along the hyperspherical manifold to synthesize a diverse set of unknown prototypes. As illustrated in Fig. 1(b), this manifold-aware expansion ensures that the generated representations remain on the valid boundary of the feature space, helping preserve their structural validity and visual objectness while effectively covering the vast and unmapped semantic regions associated with OOD targets.

Furthermore, we introduce a dual-constraint selection mechanism to strike a delicate balance between maintaining known-class discrimination and enhancing unknown-target recall. We enforce *Geodesic Pruning* as a strict geometric hard constraint to ensure that the newly traversed prototypes maintain a safe metric margin from all established class boundaries, preventing manifold collisions. This is complemented by *Text-Guided Semantic Steering*, which leverages domain-level semantic attractors—automatically inferred from known category semantics via a Large Language Model (LLM) to ensure strict adherence to the OWOD task constraints—to rescue candidate prototypes from semantically dormant regions of the manifold.

In summary, the main contributions of this paper are:

- We identify a critical geometric bottleneck in VLM-based OWOD, highlighting how Euclidean averaging of known features on a hyperspherical manifold leads to mean collapse, thereby degrading the visual objectness of unknown representations.
- We propose MGME, a framework that leverages Riemannian geometry to synthesize robust unknown representations. By performing orthogonal geodesic exploration, we generate diverse multi-prototypes that remain on the valid manifold boundary.
- We design a dual-constraint selection mechanism that unifies geodesic pruning with text-guided steering to ensure the generated prototypes are both safe and semantically aligned.
- Extensive experiments on the Real-World Object Detection (RWD) benchmark demonstrate that MGME establishes a new state-of-the-art, significantly enhancing unknown detection performance while maintaining high stability for established known classes.

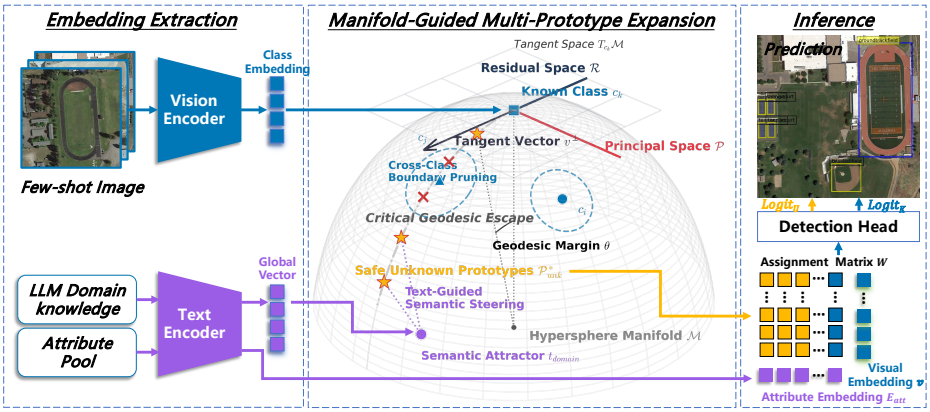

Figure 2: Overall pipeline of the proposed Manifold-Guided Multi-Prototype Expansion framework.

2 Related Work

2.1 Open World Object Detection with Foundation Models

Open World Object Detection (OWOD) extends standard closed-set detection paradigms by requiring models to simultaneously identify known classes and discover previously unseen objects without any prior category information Joseph et al. (2021); Gupta et al. (2022); Zohar et al. (2023b); Zhao et al. (2023); Li et al. (2024b); Liang et al. (2023); Liu et al. (2025). Recently, the integration of multi-modal foundation models has significantly advanced Open-Vocabulary Object Detection (OVOD), leveraging rich vision-language alignment for remarkable zero-shot generalization to novel concepts Zareian et al. (2021); Gu et al. (2021); Zhong et al. (2022); Minderer et al. (2022); Kuo et al. (2022); Yao et al. (2023); Maaz et al. (2022). However, while OVOD models excel at localizing novel targets, they fundamentally rely on explicit text prompts, such as predefined category names. This reliance makes them inherently unsuitable for the strict OWOD setting, where the system must discover true unknowns without any given information about the unseen categories Joseph et al. (2021); Xi et al. (2024).

To address this limitation, recent state-of-the-art OWOD frameworks have utilized foundation models and attribute-driven paradigms to model the unknown space. Approaches such as FOMO Zohar et al. (2023a), UMB Xi et al. (2024), and PASS Yang et al. (2025) decompose abstract categories into fine-grained, class-agnostic attributes, enabling the detection of novel objects that share visual or functional characteristics with known classes. Despite these advancements, existing methods predominantly rely on implicit Euclidean assumptions for unknown modeling, which leads to manifold collapse and semantically ambiguous representations. To address this, we adopt a Riemannian perspective to synthesize geometrically valid unknown prototypes via geodesic exploration within a focused orthogonal residual subspace.

2.2 Representation Learning on Riemannian Manifolds

Traditional deep learning architectures predominantly operate under the assumption that data is embedded in a flat Euclidean space. However, real-world data representations, particularly those bridging visual and semantic domains, frequently exhibit complex hierarchical structures and highly anisotropic distributions. To more accurately model these intrinsic topologies, representation learning on Riemannian manifolds—such as hyperbolic spaces and hyperspheres—has gained significant traction Khrulkov et al. (2020); Liu et al. (2020); Chami et al. (2019); Du et al. (2022); Li et al. (2024a). By leveraging non-Euclidean geometries, these models offer richer representation capacities and better adaptability to open-environment tasks where uniform distribution assumptions fail.

Generating new feature samples or augmenting data within these curved spaces presents unique challenges. Applying standard Euclidean operations, such as linear interpolation or arithmetic averaging, to manifolds with non-zero curvature inevitably leads to severe geometric distortion and topological collapse. To address this, recent advancements in manifold-based feature generation

utilize intrinsic geometric metrics to synthesize valid samples. For instance, methods like R-Mixup Kan et al. (2023) introduce log-Euclidean distances to prevent the inflation effects caused by flat interpolation, while HypMix Sawhney et al. (2021) and distribution-based hyperbolic augmentation Gao et al. (2022) employ manifold-specific mathematical operations to generate robust variations without distorting the underlying geometric structure. Inspired by these principles, our framework synthesizes unknown representations via geodesic traversal within the orthogonal residual subspace, ensuring visual objectness and structural validity while circumventing the geometric distortions inherent in Euclidean aggregation.

3 METHODOLOGY

We introduce MGME for OWOD, as illustrated in Fig. 2. We first formulate unknown modeling from a Riemannian perspective, then describe orthogonal residual exploration, geodesic prototype generation, semantic steering, and inference integration.

3.1 MOTIVATION AND PROBLEM FORMULATION

In OWOD, a detector must recognize K known classes while simultaneously identifying instances of unknown categories. A core challenge lies in constructing an effective representation for these unseen objects without bounding box annotations. Existing approaches typically compute the arithmetic mean of known class embeddings, $\mu_{unk} = \frac{1}{K}\sum_{k=1}^{K} c_k$, or utilize a single text prompt to serve as a universal proxy for the unknown.

This conventional paradigm suffers from two fundamental flaws. Conceptually, a single mean embedding inherently treats the "unknown" as a monolithic "objectness" prior. However, real-world unknown targets are distinct, out-of-distribution (OOD) novel entities. Representing this unbounded diversity with a single point estimate results in a near-zero probability measure for activating diverse features.

Geometrically, existing methods implicitly operate under a flat Euclidean assumption. In modern Vision-Language Models (VLMs), feature embeddings are strictly L_2-normalized for cross-modal alignment, constraining the feature space to a $(d-1)$-dimensional hyperspherical Riemannian manifold, $\mathcal{M} = \mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}$. Even after re-normalization, such a single mean proxy still compresses diverse semantic directions into one representation, limiting its ability to capture the multimodal diversity of unknown objects.

To circumvent mean collapse, we model the unknown as a diverse set of multi-prototypes distributed along the valid boundaries of the known classes. Formally, let $\mathcal{K} = \{c_1, \ldots, c_K\} \subset \mathcal{M}$ denote the set of known class centers. Our objective is to synthesize a robust set of unknown prototypes $\mathcal{P}_{unk} = \{\tilde{p}_1, \ldots, \tilde{p}_U\} \subset \mathcal{M}$ covering the OOD semantic space. For each known center c_k, we aim to identify a semantic exploration direction $v^{\perp}$ residing in its tangent space $T_{c_k}\mathcal{M}$. The generation of a valid unknown prototype $\tilde{p}_{unk}$ is mathematically formulated as a geodesic traversal:

$$\tilde{p}_{unk} = \exp_{c_k}(\theta v^{\perp}), \quad \text{s.t. } v^{\perp} \in \mathcal{R}_k \subset T_{c_k}\mathcal{M}, \tag{1}$$

where $\exp_{c_k}(\cdot)$ denotes the Riemannian exponential map ensuring $\tilde{p}_{unk} \in \mathcal{M}$, $\mathcal{R}_k$ is the orthogonal residual subspace with reduced alignment to the dominant known-class variance, and θ serves as a critical geodesic margin separating novel concepts from known distributions.

3.2 EXPLORATION IN ORTHOGONAL RESIDUAL SUBSPACE

To solve the geodesic traversal defined in Eq. 1, the first critical step is to construct a valid exploration subspace that guarantees both semantic novelty and visual objectness. The known concepts do not uniformly occupy the high-dimensional VLM feature space; rather, they densely cluster on a lower-dimensional sub-manifold Jing et al. (2021).

Intrinsic Manifold Extraction. We first estimate the intrinsic dimensionality of the known world. Applying Eigenvalue Decomposition (EVD) to the empirical covariance matrix of all L_2-normalized known instances yields eigenvectors $Q = [q_1, \ldots, q_d]$ and descending eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$. The principal space $\mathcal{P}$, which captures the dominant known semantics, is spanned by the top r

eigenvectors explaining a variance threshold τ_{var} (detailed in Appendix). Any significant perturbation within $\mathcal{P}$ would likely result in manifold collisions with existing known classes.

Consequently, the ideal domain for OOD exploration lies in the orthogonal complement of $\mathcal{P}$, referred to as the residual subspace. However, exploring the extreme tail-end eigenvectors ($\lambda_i \to 0$) presents a critical hazard. As demonstrated by prior studies on dimensionality collapse, VLM embeddings concentrate meaningful variations solely within dominant dimensions; extreme residual dimensions predominantly encode network biases or non-semantic noise. Injecting arbitrary variance into these dormant dimensions generates features completely outside the natural image manifold, stripping the resulting prototype of essential visual "objectness".

To circumvent this trap, we extract a restricted continuous band of D_{res} eigenvectors immediately following the principal boundary r. We formulate the focused orthogonal residual subspace basis $V_{res} = [q_{r+1}, \ldots, q_{r+D_{res}}] \in \mathbb{R}^{d \times D_{res}}$. This basis is strictly orthogonal to $\mathcal{P}$ (reducing alignment with the dominant known semantic directions) and explicitly filters out extreme bottom-tier dimensions.

Tangent Space Orthogonalization. To approximate the diverse distribution of the open world, we generate an initial perturbation vector by performing isotropic sampling within the truncated residual subspace: $z \sim \mathcal{N}(0, I_{D_{res}})$, which is then projected back to the ambient space via $v_{res} = V_{res} z$.

While v_{res} is geometrically orthogonal to the global principal space $\mathcal{P}$, any valid movement originating from a specific class center c_k must strictly lie within its local tangent space $T_{c_k}\mathcal{M}$ (i.e., $\langle v, c_k \rangle = 0$). To rigorously satisfy this topological constraint and prevent semantic leakage, we perform an explicit Gram-Schmidt orthogonalization. We project v_{res} onto $T_{c_k}\mathcal{M}$ by subtracting its parallel component and L_2-normalize it to obtain a unit tangent vector:

$$v^{\perp} = \frac{v_{res} - \langle v_{res}, c_k \rangle c_k}{\| v_{res} - \langle v_{res}, c_k \rangle c_k \|_2}. \tag{2}$$

This pure, linearly independent velocity vector $v^{\perp} \in T_{c_k}\mathcal{M}$ prepares us for safely pushing the representation out of the known boundaries.

3.3 Critical Geodesic Escape

Equipped with the unit tangent vector $v^{\perp} \in T_{c_k}\mathcal{M}$ providing a collision-free exploration direction, the remaining objective is to safely traverse the manifold to synthesize the unknown prototype $\tilde{p}_{unk}$.

The magnitude of this traversal dictates a strict trade-off: insufficient displacement fails to escape the known attractive basin, causing manifold collisions; whereas excessive displacement pushes the feature into semantic dead zones, stripping away visual "objectness". To resolve this, we identify a *key geodesic margin* θ to optimally breach the known decision boundary while preserving structural validity.

Valid traversal on the hypersphere $\mathbb{S}^{d-1}$ requires mapping the tangent vector via the Riemannian exponential map. Given the orthogonal condition $\langle c_k, v^{\perp} \rangle = 0$, the geodesic traversal of arc length θ from c_k yields a precise analytical form:

$$\tilde{p}_{unk} = \exp_{c_k}(\theta v^{\perp}) = \cos(\theta) c_k + \sin(\theta) v^{\perp}. \tag{3}$$

This exponential map inherently guarantees that $\tilde{p}_{unk} \in \mathbb{S}^{d-1}$. Semantically, the component $\cos(\theta) c_k$ acts as a structural anchor to retain valid physical traits, while $\sin(\theta) v^{\perp}$ injects orthogonal novelty.

By evaluating a set of discrete values for θ, we generate a candidate pool of virtual prototypes at varying depths of the residual subspace. These candidates are subsequently filtered via topology pruning (Sec. 3.4) to ensure boundary safety.

3.4 Dual-Constraint Selection

While the tangent direction $v^{\perp}$ is strictly orthogonal to its parent class center c_k, the generated candidate $\tilde{p}_{unk}$ retains a controlled similarity to c_k and may inadvertently encroach upon the decision boundaries of other known classes c_j ($j \neq k$). Furthermore, isotropic exploration in the residual

subspace may yield mathematically valid yet semantically meaningless directions (e.g., pure background textures). To bridge this semantic gap while preventing catastrophic manifold collisions, we formulate prototype selection as a constrained optimization problem, unifying a geometric hard constraint with a cross-modal soft ranking.

Cross-Class Boundary Pruning. To reduce cross-class interference, we enforce a geometric similarity constraint relative to every non-parent known class center. A candidate prototype $\tilde{p}$ is retained if and only if its maximum cosine similarity to any other known class center falls below a predefined safety margin τ_{safe}. This dynamic pruning establishes a filtered candidate pool $\mathcal{C}_{safe}$:

$$\mathcal{C}_{safe} = \left\{ \tilde{p} \;\middle|\; \max_{j \neq k} \langle \tilde{p}, c_j \rangle < \tau_{safe} \right\}. \tag{4}$$

If a local region is topologically deadlocked by surrounding known classes (i.e., $\mathcal{C}_{safe} = \emptyset$), the algorithm refrains from generating a prototype to strictly prioritize the preservation of established classification boundaries.

Text-Guided Semantic Steering. To rescue surviving candidates from semantic dead zones, we introduce a cross-modal steering mechanism. Since specifying exact unknown class names violates the strict zero-shot assumption of OWOD, we leverage Large Language Models (LLMs) as a domain knowledge prior. Given the known categories, we prompt an LLM to hallucinate a diverse list of potential unknown objects likely to co-occur in the specific domain. We extract their text embeddings via the VLM's text encoder and compute their global mean vector t_{domain}. This mean vector smooths out individual hallucination variance and acts as a robust, domain-level *semantic attractor*.

From the topologically safe pool $\mathcal{C}_{safe}$, we select the optimal unknown prototype $\tilde{p}^*_{unk}$ that maximizes alignment with the semantic attractor t_{domain}:

$$\tilde{p}^*_{unk} = \arg \max_{\tilde{p} \in \mathcal{C}_{safe}} \langle \tilde{p}, t_{domain} \rangle. \tag{5}$$

3.5 Training Adaptation and Inference

Having established the multi-prototype generation mechanism, we now detail the integration of these prototypes into the training and inference pipeline. We first introduce an orthogonal regularization to consolidate the known semantic boundaries, followed by the augmented inference protocol.

Discriminative Prototype Orthogonalization. While our manifold-guided expansion strategy effectively explores the residual subspace, maintaining sharp discriminative margins among *known* categories remains a prerequisite. Distinct categories often share common attribute descriptions, leading to entangled text-derived classifiers. During training adaptation, the model constructs known text prototypes by optimizing the fine-grained attribute embeddings $E_{att} \in \mathbb{R}^{M \times d}$ and the assignment matrix $W \in \mathbb{R}^{M \times K}$. The assembled text prototypes, $P_{cls} = W^\top E_{att} \in \mathbb{R}^{K \times d}$, are pulled towards the *frozen* visual class centers via a base adaptation loss $\mathcal{L}_{base}$ (detailed in Appendix).

To prevent semantic crowding in dense manifold regions, we introduce a mutual orthogonality constraint directly on the assembled text prototypes. Treating P_{cls} as points on the hypersphere, we L_2-normalize them to $\tilde{P}_{cls} = P_{cls}/\|P_{cls}\|_2$ and define the discriminative orthogonal loss:

$$\mathcal{L}_{ortho} = \left\| \tilde{P}_{cls} \tilde{P}_{cls}^\top - I \right\|_F. \tag{6}$$

By penalizing high off-diagonal values, $\mathcal{L}_{ortho}$ dynamically repels the text prototypes during optimization. This regularization forces the network to discover mutually exclusive attribute combinations, sharpening the decision boundaries. The final training objective is $\mathcal{L}_{total} = \mathcal{L}_{base} + \lambda_{ortho}\mathcal{L}_{ortho}$. Ultimately, consolidating the known semantic space minimizes inter-class confusion and establishes a well-separated geometric foundation for the unknown prototypes.

Multi-Prototype Inference. During inference, the diverse set of U collision-free unknown prototypes (Sec. 3.4) is integrated into the detection head. We augment the assignment matrix to $W \in \mathbb{R}^{M \times (K+U)}$, accommodating both known and unknown representations. For a visual embedding v, the unified raw logits are computed via cosine similarity:

$$Logit_{raw} = \cos(v, E_{att}) \cdot W = [Logit_K, Logit_{U_1}, \ldots, Logit_{U_U}]. \tag{7}$$

Table 1: OWOD results on five real-world datasets. We report U/K-mAP for Task 1 and PK/CK-mAP for Task 2. Bold values denote the best results, and Δ indicates the gain over the strongest baseline. For fair comparison, UMB and PASS results are reproduced under the same OWL-ViT pipeline; therefore, the values may differ from those reported in the original papers (Appendix C.3).

Dataset (→)	**Aquatic**				**Aerial**				**Game**				**Medical**				**Surgery**				**Overall**			
Task ID (→)	**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	45.1	36.5	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
FOMO Zohar et al. (2023a)	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
UMB Xi et al. (2024)	9.2	43.7	43.5	41.1	1.6	18.8	19.2	5.2	13.9	3.5	4.6	4.1	0.9	5.4	6.5	3.2	7.7	13.5	14.1	10.7	6.7	17.0	17.6	12.9
PASS Yang et al. (2025)	5.4	43.1	43.2	50.5	1.8	17.4	18.0	6.8	15.1	8.3	5.2	5.6	1.1	8.3	7.0	4.4	10.8	14.6	15.1	18.2	6.8	18.3	17.7	17.1
MGME (Ours)	9.4	44.9	44.9	53.7	2.1	18.9	21.4	6.8	22.5	9.4	6.2	7.3	9.1	8.1	8.0	3.1	15.8	14.6	11.9	23.5	**11.8**	**19.2**	**18.5**	**18.9**
Δ	**+0.2**	**+1.1**	**+0.8**	**+3.2**	**+0.3**	**+0.1**	**+2.2**	**0.0**	**+7.4**	**+1.1**	**+1.0**	**+1.7**	**+7.0**	**-0.2**	**+1.0**	**-8.4**	**+5.0**	**0.0**	**-3.2**	**+5.3**	**+5.0**	**+0.9**	**+0.8**	**+1.8**
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	0.9	0.1	0.1	0.2	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
FOMO Zohar et al. (2023a)	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
UMB Xi et al. (2024)	19.1	49.3	51.0	49.7	6.5	39.6	37.9	22.4	34.3	10.8	10.3	12.1	3.9	21.8	19.0	34.1	15.9	39.6	39.7	17.9	15.9	32.2	31.6	27.2
PASS Yang et al. (2025)	15.1	53.5	54.5	56.6	4.6	35.9	35.7	16.8	33.2	22.6	21.6	21.2	6.6	27.6	27.1	30.3	16.6	44.7	44.4	33.7	15.2	36.9	36.7	31.7
MGME (Ours)	20.6	52.8	56.4	59.5	12.6	42.0	39.5	24.6	34.4	22.0	19.6	23.8	15.1	28.3	26.9	32.5	13.5	45.5	43.3	37.8	**19.2**	**38.1**	**37.1**	**35.6**
Δ	**+1.5**	**-0.7**	**+1.9**	**+2.9**	**+6.1**	**+2.4**	**+1.6**	**+2.2**	**+0.1**	**-0.6**	**-2.0**	**+2.6**	**+5.7**	**+0.7**	**-0.2**	**-2.1**	**-3.1**	**+0.8**	**-1.1**	**+4.1**	**+3.3**	**+1.2**	**+0.4**	**+3.9**

Because a genuine unknown object typically occupies a localized region within the residual subspace, it strongly aligns with a specific prototype rather than the global distribution. Therefore, we apply a max-pooling operation to derive the definitive unknown score: $Logit_U^* = \max_i(Logit_{U_i})$.

Following standard baseline protocols, the final probability for known classes is $P_K(v) = \sigma(Logit_K)$. For the unknown targets, the base score is modulated by attribute relevance p_{ID} and outlier suppression p_{OOD} (formulations detailed in Appendix):

$$P_U(v) = \sigma(Logit_U^*) \cdot p_{ID} \cdot p_{OOD}. \tag{8}$$

By replacing a weak monolithic proxy with the robust $Logit_U^*$ derived from the multi-prototype representation, our framework significantly enhances open world sensitivity without compromising established known-class boundaries.

4 Experiments

4.1 Experimental Setup

We evaluate MGME on the Real-World Object Detection (RWD) benchmark Zohar et al. (2023a), which encompasses five diverse domains: **Aquatic** (underwater), **Aerial** (drone-captured), **Game** (synthetic), **Medical** (X-ray), and **Surgery** (surgical video). Following the standard OWOD protocol, we report Known (K) and Unknown (U) mAP for Task 1, and Previously Known (PK) and Currently Known (CK) mAP for Task 2. Our framework, implemented in PyTorch on NVIDIA RTX 4090 GPUs using the OWL-ViT Minderer et al. (2022) backbone, is compared against SOTA methods including FOMO Zohar et al. (2023a), UMB Xi et al. (2024), and PASS Yang et al. (2025). We also consider zero-shot baselines: **BASE-ZS** (generic prompts), **BASE-ZS+LLM** (using the identical LLM-generated vocabulary as our semantic attractor for fair comparison), and **BASE-ZS+GT** (ground-truth names as an upper bound). All detailed hyperparameter configurations (e.g., $\tau_{safe}, \theta, \lambda_{ortho}$) and implementation details are provided in Appendix.

4.2 Comparison with State-of-the-Arts

Table 1 presents the comprehensive quantitative results across the five datasets of the RWD benchmark. Overall, MGME achieves state-of-the-art performance across diverse domains, significantly advancing the boundaries of OWOD. Under the ViT-B/16 backbone, MGME achieves an impressive

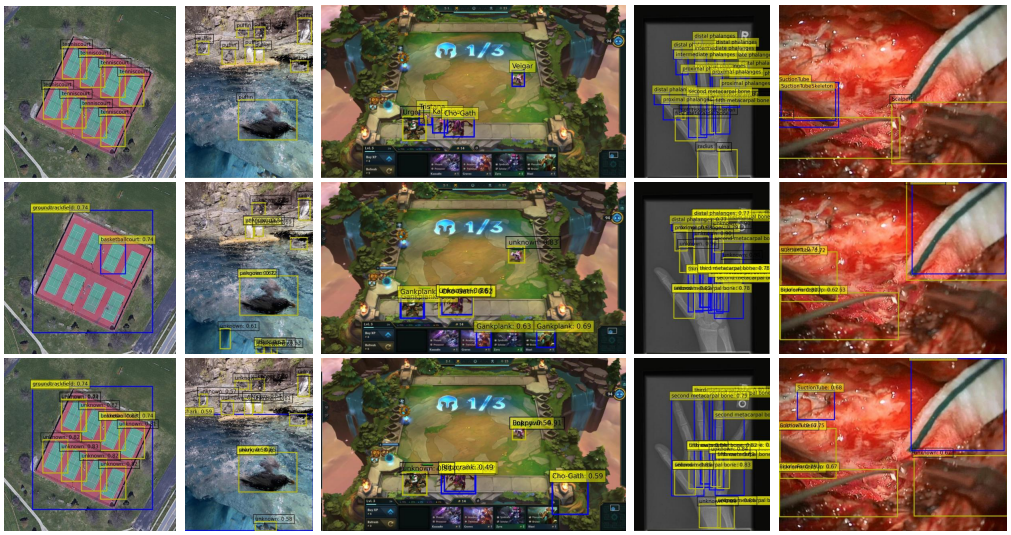

Figure 3: **Qualitative comparison of OWOD results on RWD datasets.** Columns correspond to Aerial, Aquatic, Game, Medical, and Surgery. Rows show Ground Truth, PASS, and MGME. Blue and yellow boxes indicate known and unknown.

overall Unknown mAP (U-mAP) of 11.8, outperforming the strongest baseline (PASS) by a remarkable absolute margin of +5.0. Concurrently, the Known mAP (K-mAP) is also improved to 19.2 (+0.9). Under the ViT-L/14 backbone, MGME achieves an overall U-mAP of 19.2 (+3.3 over UMB) and a K-mAP of 38.1 (+1.2 over PASS). The improvements in both metrics validate our hypothesis: geometrically sound unknown exploration can significantly improve unknown performance without degrading known-class accuracy.

The performance gains of MGME are consistent with the benefits of its manifold-guided multi-prototype design. Zero-shot baselines (e.g., BASE-ZS) assume a monolithic "objectness" prior, yielding a low U-mAP (e.g., 4.9 on B/16) because real-world novel targets are diverse out-of-distribution (OOD) entities. Notably, even when equipping the baseline with the exact same LLM-hallucinated vocabulary used in our method (BASE-ZS+LLM), the overall U-mAP reaches 10.2. It still falls short of our method, demonstrating that simply injecting semantic text is insufficient without rigorous geometric alignment. Furthermore, recent few-shot methods (FOMO, UMB, PASS) rely on the arithmetic mean of known embeddings to anchor the unknown space. As theorized in Sec. 3.1, Euclidean averaging in the highly anisotropic VLM space may mathematically cause features to deviate from the valid data manifold into semantically ambiguous regions. Moreover, a single mean center cannot represent the diversity of open world distributions.

Our multi-prototype geodesic escape effectively mitigates this issue, with performance gains being particularly notable in specific domains. In the Medical dataset (B/16), PASS achieves 1.1 U-mAP, whereas MGME reaches 9.1 (+8.0). Similarly, in the Game dataset (B/16), MGME reaches 22.5 U-mAP (+7.4 over PASS). Fine-grained features in these domains reside in extremely dense submanifolds, where mean interpolation may destroy the visual "objectness" required for detection. By performing orthogonal traversal in the residual subspace, MGME successfully synthesizes distinct prototypes that capture subtle OOD variations. In the Aerial dataset (L/14), MGME not only nearly doubles the U-mAP from 6.5 (UMB) to 12.6 (+6.1), but the K-mAP also simultaneously improves from 39.6 to 42.0 (+2.4). This suggests that our dynamic cross-class boundary pruning helps reduce manifold collisions, limiting the interference of newly generated prototypes with established classes and helping refine the overall decision boundaries. Despite overall superiority, minor degradations occur in extreme scenarios like Surgery U-mAP (-3.1 on L/14). When unknown targets share overwhelming similarities with known classes, the valid residual subspace becomes exceedingly narrow. Satisfying the safety constraint forces a larger traversal margin θ, inadvertently pushing prototypes away from the closely entangled unknowns, which slightly reduces their recall.

Table 2: Ablation study of MGME. We report U/K-mAP on Aerial and Medical datasets.

Dataset (→)	**Aerial**		**Medical**	
	U	K	U	K
MGME Full (Ours)	**12.6**	**42.0**	**15.1**	**28.3**
w/o Subspace Truncation (D_{res})	11.8	42.0	10.0	24.4
w/o Geodesic Escape (Euclidean Add.)	5.6	39.9	1.8	28.2
w/o Boundary Pruning (τ_{safe})	10.7	42.1	15.1	25.3
w/o Semantic Steering (t_{domain})	10.2	42.1	14.8	27.0
w/o Orthogonalization ($\mathcal{L}_{ortho}$)	10.2	40.9	15.6	24.8

4.3 Ablation Study

To validate the geometric designs of MGME, we conduct ablation studies on the Aerial and Medical datasets using the ViT-L/14 backbone (Table 2). We first replace the geodesic escape in Eq. 3 with the Euclidean-addition variant used in our implementation. This change causes a substantial drop in U-mAP (12.6 to 5.6 on Aerial; 15.1 to 1.8 on Medical), showing that the proposed geodesic construction is more effective than this particular Euclidean-addition variant under the current pipeline. Similarly, expanding across the full residual subspace without truncation (D_{res}) degrades both U-mAP and K-mAP (e.g., Medical U-mAP drops to 10.0), validating that extreme tail-end eigenvectors predominantly encode non-semantic noise that interferes with novelty discovery.

Furthermore, we evaluate the mechanisms for ensuring exploration safety and semantic alignment. Disabling the topological hard constraint (τ_{safe}) leads to a significant K-mAP drop on Medical (28.3 to 25.3) and U-mAP drop on Aerial (12.6 to 10.7), suggesting that unrestrained expansion can increase manifold collisions and inter-class confusion. Removing the LLM-guided semantic steering (t_{domain}) results in an Aerial U-mAP of 10.2, confirming that unguided geometric exploration may inadvertently point toward background artifacts. Finally, ablating the discriminative orthogonalization loss ($\mathcal{L}_{ortho}$) during adaptation impairs known-class discriminability, with K-mAP falling to 40.9 and 24.8 on Aerial and Medical, respectively. This underscores the importance of penalizing feature entanglement to establish a clean geometric foundation for open world expansion.

4.4 Qualitative Results

Fig. 3 compares the qualitative detection results of MGME and PASS. In fine-grained domains, representing unknowns with a single mean embedding causes features to collapse into semantic dead zones. In Medical, PASS fails to detect unknown bones (Radius and Ulna), whereas MGME generates diverse prototypes that capture these subtle OOD variations. Similarly, PASS misses the unknown Scalpel in Surgery and several unknown Puffins in Aquatic, while MGME successfully discovers and localizes them.

In highly entangled feature spaces, baseline methods suffer from manifold collisions and known–unknown confusion. In the Aerial dataset, the scene entirely consists of unknown Tennis courts that closely resemble known Basketball courts. PASS misses a significant portion of these courts and erroneously classifies the detected ones as known Basketball courts. In contrast, MGME fully recalls all Tennis courts and correctly identifies them as unknown entities. In the Game dataset, PASS produces false unknown predictions due to inter-class confusion, whereas MGME mitigates this issue through cross-class boundary pruning and semantic steering. This design preserves established known-class margins while enabling accurate discovery of novel entities.

5 Conclusion

MGME addresses mean collapse in OWOD by generating diverse unknown prototypes via orthogonal geodesic traversal. Its dual-constraint selection combines boundary pruning and semantic steering to preserve known-class margins while aligning novel prototypes with domain semantics, while prototype orthogonalization further improves discrimination. Experiments on RWD demonstrate stronger unknown discovery without sacrificing known-class performance.

AI USE STATEMENT

Generative AI tools were used for research execution to generate candidate unknown-class concepts for the semantic-steering component of MGME and to generate the initial attribute pools used by the attribute-based baselines. The generated concept lists used in the experiments were fixed before evaluation and were applied consistently across comparisons. Generative AI tools were also used to aid language polishing of the manuscript. All AI-assisted outputs were reviewed by the authors, and the experimental results and claims reported in this paper were verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies open-world object detection using existing datasets from the RWD benchmark and does not introduce a new dataset. The benchmark includes medical, surgical, and aerial domains, where false positive or false negative predictions may have significant consequences in real-world deployment. The results reported in this work are research evaluations and should not be interpreted as evidence that MGME is suitable for autonomous high-stakes decision making. In addition, the semantic-steering component relies on LLM-generated concepts, which may vary across model versions and prompts and may inherit biases from the underlying language model. We mitigate this source of variability in our experiments by using a fixed, pre-generated concept list. Further domain-specific validation, safeguards, and human oversight would be necessary before deployment in safety-critical applications.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate reproducibility in the Appendix, including the mathematical formulations of MGME, complete algorithm pseudo-code, dataset and evaluation protocols, baseline implementation details, hyperparameter configurations, parameter-sensitivity analysis, and additional qualitative results. All compared methods are evaluated under a unified OWL-ViT pipeline using consistent RWD splits and evaluation code. The Appendix also documents the use of LLM-generated concepts and the settings used for the proposed geometric expansion and pruning mechanisms.

REFERENCES

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022.

Zhi Gao, Yuwei Wu, Yunde Jia, and Mehrtash Harandi. Hyperbolic feature augmentation via distribution estimation and infinite sampling on manifolds. *Advances in neural information processing systems*, 35:34421–34435, 2022.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9235–9244, 2022.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5830–5840, 2021.

Xuan Kan, Zimu Li, Hejie Cui, Yue Yu, Ran Xu, Shaojun Yu, Zilong Zhang, Ying Guo, and Carl Yang. R-mixup: Riemannian mixup for biological networks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1073–1085, 2023.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6418–6428, 2020.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022.

Jeng-Lin Li, Ming-Ching Chang, and Wei-Chao Chen. Learning multi-manifold embedding for out-of-distribution detection. In *European Conference on Computer Vision*, pp. 417–433. Springer, 2024a.

Yiming Li, Yi Wang, Wenqian Wang, Dan Lin, Bingbing Li, and Kim-Hui Yap. Open world object detection: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2): 988–1008, 2024b.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Wenteng Liang, Feng Xue, Yihao Liu, Guofeng Zhong, and Anlong Ming. Unknown sniffer for object detection: Don't turn a blind eye to unknown objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3230–3239, 2023.

Haomiao Liu, Hao Xu, Chuhuai Yue, and Bo Ma. Un-detr: Promoting objectness learning via joint supervision for unknown object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5442–5450, 2025.

Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. Hyperbolic visual embedding learning for zero-shot recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9273–9281, 2020.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *European conference on computer vision*, pp. 512–531. Springer, 2022.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pp. 728–755. Springer, 2022.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ramit Sawhney, Megh Thakkar, Shivam Agarwal, Di Jin, Diyi Yang, and Lucie Flek. Hypmix: hyperbolic interpolative data augmentation. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 9858–9868, 2021.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in neural information processing systems*, 34:144–157, 2021.

Xing Xi, Yangyang Huang, Zhijie Zhong, and Ronghua Luo. Umb: Understanding model behavior for open-world object detection. *Advances in Neural Information Processing Systems*, 37:74233–74261, 2024.

Muli Yang, Gabriel James Goenawan, Huaiyuan Qin, Kai Han, Xi Peng, Yanhua Yang, and Hongyuan Zhu. Detecting open world objects via partial attribute assignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20318–20328, 2025.

Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23497–23506, 2023.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14393–14402, 2021.

Xiaowei Zhao, Yuqing Ma, Duorui Wang, Yifan Shen, Yixuan Qiao, and Xianglong Liu. Revisiting open world object detection. *IEEE transactions on circuits and systems for video technology*, 34 (5):3496–3509, 2023.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16793–16803, 2022.

Orr Zohar, Alejandro Lozano, Shelly Goel, Serena Yeung, and Kuan-Chieh Wang. Open world object detection in the era of foundation models. *arXiv preprint arXiv:2312.05745*, 2023a.

Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11444–11453, 2023b.

A Supplementary Mathematical Formulations

This section provides the detailed mathematical formulations and baseline configurations omitted from the main text due to space constraints.

A.1 Cumulative Variance and Intrinsic Dimensionality

As introduced in Sec. 3.2 of the main paper, we estimate the intrinsic dimensionality of the known class feature distributions to delineate the semantic boundaries of the established categories. Given the empirical covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ of the L_2-normalized known features, we perform Eigenvalue Decomposition (EVD) to obtain $\Sigma = Q\Lambda Q^\top$. The diagonal matrix $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ contains the eigenvalues sorted in descending order ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$), which represent the magnitude of data variance captured along each corresponding orthogonal eigenvector in Q.

To quantify the semantic capacity of these principal directions, we define the cumulative variance ratio for the top m eigenvectors as:

$$C(m) = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{d} \lambda_j}. \tag{9}$$

The intrinsic dimensionality r of the principal space $\mathcal{P}$ is thereby determined by the minimum number of components required to reach a predefined variance threshold τ_{var}:

$$r = \min \{m \mid C(m) \geq \tau_{var}\}. \tag{10}$$

This truncation strategy is deeply rooted in the geometric properties of VLM embedding spaces. Due to the inherent dimensionality collapse during large-scale contrastive pre-training, VLMs produce highly anisotropic distributions. Consequently, the vast majority of semantically meaningful variations (e.g., visual attributes defining base categories such as shape, texture, and color) are densely concentrated within a remarkably small subset of dominant dimensions. By identifying the boundary r that encapsulates τ_{var} of the total variance, we effectively map out the structural semantics of

the known world. This guarantees that subsequent orthogonal explorations in the residual subspace do not inadvertently alter these core foundational attributes.

The precise value of τ_{var} dictates the strictness of this semantic boundary. A higher threshold preserves more known variations but narrows the available residual subspace for exploration, whereas a lower threshold expands the novelty capacity at the risk of semantic leakage and manifold collisions.

A.2 Detailed Base Adaptation Loss

As introduced in Sec. 3.5 of the main paper, our framework builds upon the established baseline by optimizing the attribute embeddings and the assignment weight matrix W during the training adaptation phase. This optimization is supervised by a composite base adaptation objective, formulated as:

$$\mathcal{L}_{base} = \mathcal{L}_{CE} + \lambda_{ot}\mathcal{L}_{ot} + \lambda_{mse}\mathcal{L}_{MSE} + \lambda_{l1}\mathcal{L}_{L1}, \tag{11}$$

where the empirical trade-off coefficients are uniformly set to $\lambda_{ot} = 5$, $\lambda_{mse} = 0.1$, and $\lambda_{l1} = 0.001$ across all experiments. The detailed formulations and rationales for each component are provided below. The base adaptation objective is kept identical for the baseline and MGME. The only additional training term in MGME is $\lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}}$. The orthogonalization term acts as a soft decorrelation regularizer rather than enforcing exact semantic orthogonality. Shared class semantics are still preserved by the base adaptation loss and the alignment with frozen visual class centers.

Partial Optimal Transport Loss ($\mathcal{L}_{ot}$): Conventional Optimal Transport (OT) relies on the strict assumption of identical total probability mass between two distributions (i.e., $\|V\|_1 = \|A\|_1$). However, this assumption is frequently violated in open-world detection scenarios, as only a small subset of the redundant, LLM-generated attribute pool A semantically aligns with the visual object embeddings V. To resolve this, $\mathcal{L}_{ot}$ implements a Partial Optimal Transport (POT) mechanism. It relaxes the equal mass assumption to transport only the in-distribution mass between V and A, while explicitly filtering out superfluous out-of-distribution (OOD) attributes. This is achieved by enforcing asymmetric marginal constraints:

$$\Pi(V, A) \triangleq \left\{ T \in \mathbb{R}_{+}^{M\times N} \;\middle|\; T\mathbf{1}_N = V,\; T^{\top}\mathbf{1}_M \leq A \right\}. \tag{12}$$

Unlike traditional POT, this formulation introduces a strict equality constraint on the visual distribution V and an inequality constraint on the attribute distribution A. The optimal transport plan is obtained by minimizing $d_{POT,\epsilon}(V, A; C) \triangleq \min_{T\in\Pi(V,A)}\langle T, C\rangle_F - \epsilon h(T)$, where $h(T)$ serves as an entropic regularizer to accelerate optimization convergence.

Cross-Entropy Loss ($\mathcal{L}_{CE}$): Following the POT-based attribute selection, a mapping matrix $W \in \mathbb{R}^{M\times K}$ is learned to assign the filtered attributes to the K known object classes. The prediction probability $p(O_k|v)$ is computed using a Softmax function applied over the similarity between the visual embeddings and the reweighted attribute representations. A standard cross-entropy loss supervises both the selected attributes and the assignment matrix:

$$\mathcal{L}_{CE} = -\frac{1}{M_{patch}} \sum_{m=1}^{M_{patch}} \sum_{k=1}^{K} l_{m,k} \log p(O_k|v_m), \tag{13}$$

where $l_{m,k}$ denotes the one-hot label vector for the visual patch v_m, and M_{patch} represents the total number of known image patches within the batch.

Mean Squared Error Loss ($\mathcal{L}_{MSE}$): To bridge the modality gap, $\mathcal{L}_{MSE}$ explicitly aligns the text-derived class prototypes ($P_{cls} = W^{\top}E_{att}$) with the average vision-derived class centers extracted from the support set. Under limited bounding-box supervision (i.e., few-shot settings), this alignment ensures stable feature refinement and strict cross-modal semantic consistency. Crucially, as detailed in the main text, it is this exact anchoring mechanism that allows our proposed Discriminative Prototype Orthogonalization ($\mathcal{L}_{ortho}$) to effectively sharpen the classification boundaries.

L1 Regularization ($\mathcal{L}_{L1}$): To prevent attribute entanglement, an L_1 penalty is applied to the attribute mapping weights W (i.e., $\|W\|_1$). This regularization induces sparsity, encouraging the network to selectively activate only the most discriminative and class-specific attributes, rather than redundantly exploiting the entire attribute pool.

A.3 Open-World Inference Modulators

As introduced in Sec. 3.5 of the main paper, the base unknown score $\sigma(Logit_U^*)$ derived from the multi-prototype representation is modulated by two critical factors during inference: attribute relevance (p_{ID}) and outlier suppression (p_{OOD}). These modulators serve as dual gating mechanisms to ensure that the detected unknown proposals are both valid visual entities and genuinely out-of-distribution (OOD) from the established categories. Both p_{ID} and p_{OOD} are inherited unchanged from the adopted baseline and are applied to all corresponding variants. MGME modifies the construction of $Logit_U^*$ and adds $\mathcal{L}_{ortho}$, but does not introduce or retune these two inference gates.

Attribute Relevance (p_{ID}): To prevent the detector from erroneously activating on pure background noise or semantically meaningless visual artifacts, we evaluate the general "objectness" of a candidate region v. Since the foundational attribute embeddings $E_{att} = [a_1, a_2, \ldots, a_M]^\top \in \mathbb{R}^{M\times d}$ encompass a comprehensive vocabulary of class-agnostic visual properties (e.g., specific shapes, textures, and materials), a valid physical entity should naturally exhibit a high affinity toward at least one of these primitive attributes. We define the attribute relevance score as the maximum Sigmoid activation across all base attributes:

$$p_{ID}(v) = \max_{n\in\{1,\ldots,M\}} \sigma(a_n^\top v). \tag{14}$$

If a region fails to trigger any fundamental visual attributes, p_{ID} approaches zero. This mechanism downweights the unknown probability for non-object background patches and other weakly attribute-aligned responses.

Outlier Suppression (p_{OOD}): While p_{ID} ensures the region contains a valid object, it is equally imperative to verify that this object does not belong to any of the K established known classes. To achieve this, we adopt the Maximum Concept Matching (MCM) mechanism. We compute the Softmax probabilities exclusively over the known class logits $Logit_K \in \mathbb{R}^K$. If the maximum probability among the known classes is exceptionally high, the region is highly likely a known instance, and its unknown confidence must be heavily penalized to prevent false positives. The outlier suppression factor is formulated as:

$$p_{OOD}(v) = 1 - \max_{k\in\{1,\ldots,K\}} \frac{\exp(Logit_{K,k})}{\sum_{j=1}^{K}\exp(Logit_{K,j})}. \tag{15}$$

This formulation reduces known–unknown confusion between the known and unknown semantic spaces. By multiplying the multi-prototype base unknown score with both p_{ID} and p_{OOD}, our inference framework suppresses a portion of false-positive responses originating from both background noise and known objects, yielding a highly robust open-world detection capability.

B Algorithm Pseudo-code

To facilitate reproducibility, we provide the complete pseudo-code for the proposed Manifold-Guided Multi-Prototype Expansion (MGME) framework in Algorithm 1.

The algorithm directly reflects the integration of the hard geometric constraints and the soft cross-modal semantic ranking. Initially, the domain-level semantic attractor t_{domain} is computed by averaging the normalized text embeddings of LLM-hallucinated potential unknown classes. For the geometric foundation, the orthogonal residual subspace basis V_{res} is dynamically extracted by bypassing the principal variance components and bounding the active focus dimensions D_{res}. Unlike a mean unknown prototype used directly as a classification vector, t_{domain} is used only to rank geometrically generated candidates. It is never inserted into the detection head as the final unknown classifier; the final unknown score is computed from the retained multi-prototype set.

During the multi-prototype generation loop, for each known class center, the algorithm systematically expands outward across a set of discrete angle steps Θ. At each angular depth, a Monte Carlo sampling process generates candidate directions within the residual subspace, which are strictly orthogonalized against the parent class via the Gram-Schmidt process. Candidates that satisfy the safety margin τ_{safe} against all other known classes are added to a safe pool. Ultimately, the candidate that maximizes the cosine similarity with the semantic attractor t_{domain} is selected as the optimal prototype for that specific shell. If a local region is topologically deadlocked and no safe

Algorithm 1 Manifold-Guided Multi-Prototype Expansion(MGME)

Require: Known class centers $\mathcal{K} = \{c_1, \ldots, c_K\}$, Manifold eigenvectors Q and eigenvalues Λ, Domain LLM prompts, Number of prototypes per class U_c, Angle steps Θ, Max attempts N_{att}, Variance threshold τ_{var}, Focus dimensions D_{res}, Safety margin τ_{safe}
Ensure: Generated unknown prototype set $\mathcal{P}_{unk}$

Extract normalized text embeddings for LLM prompts and compute global mean t_{domain}
Calculate intrinsic dimensionality r using Λ and cumulative variance threshold τ_{var}
Extract the focused residual subspace basis $V_{res} = [q_{r+1}, \ldots, q_{r+D_{res}}]$
Initialize $\mathcal{P}_{unk} \leftarrow \emptyset$
for each known class $c_k \in \mathcal{K}$ **do**
  for $u = 1$ **to** U_c **do**
    $p_{opt} \leftarrow$ NULL, $s_{opt} \leftarrow -1.0$, $safe_found \leftarrow$ FALSE
    for each $\theta \in \Theta$ **do**
      for $i = 1$ **to** N_{att} **do**
        Sample isotropic noise $z \sim \mathcal{N}(0, I_{D_{res}})$
        Project to residual subspace: $v_{res} \leftarrow V_{res} z$
        Orthogonalize: $v^{\perp} \leftarrow \frac{v_{res} - \langle v_{res}, c_k \rangle c_k}{\| v_{res} - \langle v_{res}, c_k \rangle c_k \|_2}$ {Gram-Schmidt}
        Geodesic traversal: $\tilde{p} \leftarrow \cos(\theta) c_k + \sin(\theta) v^{\perp}$
        Max inter-class similarity: $s_{col} \leftarrow \max_{j \neq k} \langle \tilde{p}, c_j \rangle$ {Collision Check}
        if $s_{col} < \tau_{safe}$ **then**
          $safe_found \leftarrow$ TRUE {Hard Constraint satisfied}
          Calculate semantic alignment: $s_{text} \leftarrow \langle \tilde{p}, t_{domain} \rangle$
          if $s_{text} > s_{opt}$ **then**
            $s_{opt} \leftarrow s_{text}$, $p_{opt} \leftarrow \tilde{p}$ {Soft Ranking}
          end if
        end if
      end for
    end for
    if $safe_found$ **and** $p_{opt} \neq$ NULL **then**
      $\mathcal{P}_{unk} \leftarrow \mathcal{P}_{unk} \cup \{p_{opt}\}$ {Dynamic Topology Pruning}
    end if
  end for
end for
return $\mathcal{P}_{unk}$

candidates are found across all attempts and angles, the dynamic pruning mechanism aborts the generation to preserve the established classification boundaries.For each known class center c_k, we attempt to generate $U_c = 3$ prototypes. For each prototype slot, candidates are sampled over all angle steps $\theta \in \Theta$, and the safe candidate with the highest cosine similarity to t_{domain} is retained. If no candidate satisfies the safety constraint across all sampled directions and angles, that slot is discarded to preserve the established classification boundaries. Consequently, the final number of retained prototypes is

$$U_{\text{tot}} = |\mathcal{P}_{\text{unk}}| \leq K U_c.$$

The symbol U used in Sec. 3.5 of the main paper refers to U_{tot}.

C Detailed Experimental Setup

In this section, we provide exhaustive details regarding the experimental environment, dataset statistics, baseline configurations, and hyperparameter settings to ensure the reproducibility of the MGME framework.

C.1 Dataset Descriptions and Statistics

The Real-World Object Detection (RWD) benchmark Zohar et al. (2023a) is specifically curated to evaluate foundation model-based detectors in out-of-distribution scenarios. Unlike traditional

benchmarks like MS-COCO, the RWD datasets feature high domain shift and specialized semantic spaces. The five datasets are summarized below:

- **Aquatic:** Contains underwater imagery with challenging lighting conditions and diverse marine species.
- **Aerial:** Comprises high-resolution drone-captured images with small-scale objects and top-down perspectives.
- **Game:** Features synthetic scenes from complex video games, exhibiting high texture variance and diverse character designs.
- **Medical:** Consists of X-ray images focusing on hand bone structures, requiring extremely fine-grained morphological discrimination.
- **Surgery:** Includes frames from surgical videos where tools and anatomical structures are highly entangled and often occluded.

Following the protocol in Zohar et al. (2023a), each dataset is partitioned into Task 1 (50% of classes as Known) and Task 2 (introducing the remaining 50% as Currently Known). Specific category names and image counts for each split are consistent with the official RWD release.

C.2 Evaluation Metrics

To evaluate the model's capability in the open-world paradigm, we employ mean Average Precision (mAP) with an IoU threshold of 0.5. Let $\mathcal{C}_K$ and $\mathcal{C}_U$ be the sets of Known and Unknown classes in Task 1, respectively.

- **U-mAP:** The primary metric for open-world discovery, calculated as the mAP over classes in $\mathcal{C}_U$ where predictions are matched against unknown ground-truth boxes.
- **K-mAP:** Measures the precision on $\mathcal{C}_K$ to ensure that unknown exploration does not degrade known-class discriminability.
- **PK-mAP & CK-mAP:** In Task 2, PK-mAP evaluates performance on previously learned classes ($\mathcal{C}_K$), while CK-mAP evaluates performance on the newly introduced classes ($\mathcal{C}_U$).

C.3 Baseline Implementation Details

All baseline methods are integrated into our unified PyTorch codebase to ensure a fair comparison.

- **SOTA Methods (FOMO, UMB, PASS):** We re-train these methods using the official hyperparameters reported in their respective papers. For UMB and PASS, which involve attribute-based optimization, we employ GPT-3.5 to generate the initial attribute pool. All FOMO, UMB, and PASS results in the main paper are reproduced in our unified OWL-ViT pipeline rather than copied from the original publications. We use identical RWD splits, support samples, backbones, attribute pools, and evaluation code across methods. Consequently, the reproduced values may differ from those originally reported under different implementation and tuning configurations.
- **Zero-Shot Baselines:** All **BASE-ZS** variants are built upon the frozen OWL-ViT backbone. **BASE-ZS+LLM** uses the same set of LLM-hallucinated class names used in our semantic attractor. **BASE-ZS+GT** serves as an oracle upper bound, assuming the names of unknown classes are known a priori.
- **Fair Comparison Protocol:** To account for variances in implementation, we apply a grid search for the learning rate $LR \in \{1e-5, 5e-5, 1e-4\}$ and the maximum number of iterations $T \in \{1, 10, 100\}$ for all baselines during the adaptation phase.

C.4 Hyperparameter Configurations

The MGME framework is implemented using PyTorch on NVIDIA GeForce RTX 4090 GPUs. To facilitate the reproducibility of our geometric expansion and safety mechanisms, we summarize the key hyperparameters and their corresponding sections in the main text in Table 3.

Table 3: Hyperparameter settings for the MGME framework.

Hyperparameter	**Symbol**	**Value**	**Ref. Section**
Cumulative variance threshold	τ_{var}	0.95	Sec. 3.2
Residual subspace dimension	D_{res}	150	Sec. 3.2
Backbone dimension (d)	d	512 / 768	Sec. 3.1
Geodesic angle steps	θ	$\{20^\circ, 24^\circ, \ldots, 40^\circ\}$	Sec. 3.3
Requested prototypes per known class	U_c	3	Sec. 3.3
Topological safety margin	τ_{safe}	0.92	Sec. 3.4
Orthogonalization loss weight	λ_{ortho}	0.5	Sec. 3.5
Optimizer	-	AdamW	-

Table 4: Sensitivity analysis on the number of generated unknown prototypes (U) per known class. We report U-mAP (U) and K-mAP (K) on the Aerial and Medical datasets (Task 1 evaluation using the ViT-L/14 backbone).

Dataset ($\rightarrow$)	**Aerial**		**Medical**	
Number of Prototypes	U	K	U	K
$U = 1$	11.3	39.6	**15.4**	25.3
$U = 3$ (Ours)	**12.6**	**42.0**	15.1	28.3
$U = 5$	12.4	41.9	14.7	27.0
$U = 10$	11.8	42.0	10.3	**28.5**

D Extended Analysis

D.1 Parameter Sensitivity Analysis

To further comprehensively understand the behavior of the proposed MGME framework, we conduct a sensitivity analysis on a critical geometric hyperparameter: the number of generated unknown prototypes per known class, denoted as U. We evaluate the performance variations on the Aerial and Medical datasets using the ViT-L/14 backbone under the Task 1 protocol. The quantitative results are presented in Table 4.

Effect of the Number of Prototypes (U). As demonstrated in Table 4, generating only a single unknown prototype ($U = 1$) yields suboptimal performance, particularly in maintaining known-class discriminability (K-mAP drops significantly to 39.6 on Aerial and 25.3 on Medical). Using a single prototype yields lower K-mAP on both datasets and lower U-mAP on Aerial, although its Medical U-mAP is slightly higher than that of $U_c = 3$. Increasing the number to three provides the best overall balance between unknown recall and known-class preservation. Further increasing U_c does not yield consistent gains and reduces Medical U-mAP substantially at $U_c = 10$, indicating that additional prototypes may become redundant or interfere in a narrow residual space.

Increasing U_c beyond three does not produce consistent gains. In particular, Medical U-mAP decreases markedly at $U_c = 10$, indicating increased redundancy or interference among prototypes in a narrow residual space. Overall, $U_c = 3$ provides the best balance between unknown recall and known-class preservation.

Initial Geodesic Angle Step (θ_{start}). We evaluate different initial angles while keeping the remaining prototype-generation settings fixed. The results are reported in Table 5.

A smaller initial angle slightly improves Medical U-mAP but substantially reduces K-mAP, suggesting greater overlap with known-class regions. A larger initial angle reduces U-mAP on both datasets. The default 20° setting therefore provides the best overall trade-off under the evaluated configuration.

Table 5: Sensitivity analysis on the initial geodesic angle step (θ_{start}). We report U-mAP (U) and K-mAP (K) on the Aerial and Medical datasets (Task 1 evaluation using the ViT-L/14 backbone).

Dataset ($\rightarrow$)	**Aerial**		**Medical**	
Initial Angle (θ_{start})	U	K	U	K
$\theta_{start} = 8°$	11.5	40.1	**16.0**	24.7
$\theta_{start} = 20°$ (Ours)	**12.6**	**42.0**	15.1	**28.3**
$\theta_{start} = 32°$	11.4	39.6	14.7	25.2

D.2 Additional Qualitative Results

To further illustrate the advantages of our proposed Manifold-Guided Multi-Prototype Expansion (MGME) framework, Fig. 4 presents an additional set of qualitative results highlighting challenging scenarios where baseline methods frequently struggle. As observed in the middle row, the baseline PASS exhibits several critical failure modes due to mean collapse and manifold collisions, which MGME (bottom row) effectively overcomes. Specifically, in the **Aerial** dataset, PASS entirely misses the unknown Baseball field and erroneously classifies another as a known Ground track field, whereas MGME successfully localizes them without inter-class confusion. In the **Aquatic** dataset, visual similarity leads PASS to misclassify the unknown Stingray as a known Shark; in contrast, MGME correctly identifies it as an unknown entity, demonstrating resilience against manifold collisions. In the **Game** dataset, the baseline fails to detect heavily overlapping unknown characters, while MGME exhibits higher sensitivity in localizing these novel targets. Finally, in both the **Medical** and **Surgery** datasets, PASS misses several fine-grained unknown objects, failing to detect fine-grained unknown objects such as the Radius bone and the Scalpel. Conversely, by performing orthogonal geodesic traversal, MGME successfully discovers the subtle OOD variations required to accurately localize these structures. These direct comparisons consistently validate the superior open-world discovery capabilities of our geometric approach.



Figure 4: **Additional qualitative comparison across the five RWD datasets.** From left to right: Aerial, Aquatic, Game, Medical, and Surgery. **Top row:** Ground Truth annotations. **Middle row:** Predictions generated by the baseline (PASS). **Bottom row:** Predictions by our proposed MGME framework. Blue boxes denote known categories, while yellow boxes denote discovered unknown categories.

D.3 Failure Cases

While the proposed MGME framework demonstrates substantial improvements in open-world discovery and boundary preservation, it still encounters challenges in certain extreme scenarios. Fig. 5

illustrates several representative failure cases across the five RWD datasets, providing deeper insights into the inherent complexities of Open World Object Detection (OWOD).

First, in crowded scenes with extreme object density and small spatial scales, the local visual features required for robust orthogonal traversal can be diluted. For instance, in the **Aerial** dataset (column 1), although MGME detects significantly more unknown Tennis courts and Baseball fields compared to the baseline, it still misses a few instances. Similarly, in the **Game** dataset (column 3), heavy occlusion and complex background rendering lead to the missed detection of a known character. Furthermore, in environments with massive intra-class visual variance, the model may struggle to perfectly align visual concepts with text-derived boundaries. In the **Aquatic** dataset (column 2), similar to the baseline, MGME occasionally misclassifies known "Fish" instances that exhibit unusual appearances as unknown entities. A similar phenomenon occurs in the **Surgery** dataset (column 5), where a known "Suction tube" is misclassified as an unknown tool, highlighting the difficulty of distinguishing highly homogeneous surgical instruments. Lastly, fine-grained domains can induce severe morphological confusion, as seen in the **Medical** dataset (column 4). While the baseline PASS completely misses the target, MGME's geometric expansion successfully *localizes* the elusive "Radius" bone. However, due to the overwhelming morphological similarity among various hand bones, the text-derived classifier erroneously maps it to a known bone category rather than identifying it as unknown. This suggests that while our method excels at discovering valid physical entities in the residual subspace, extremely fine-grained semantic disambiguation remains an open challenge for future VLM-based OWOD research.

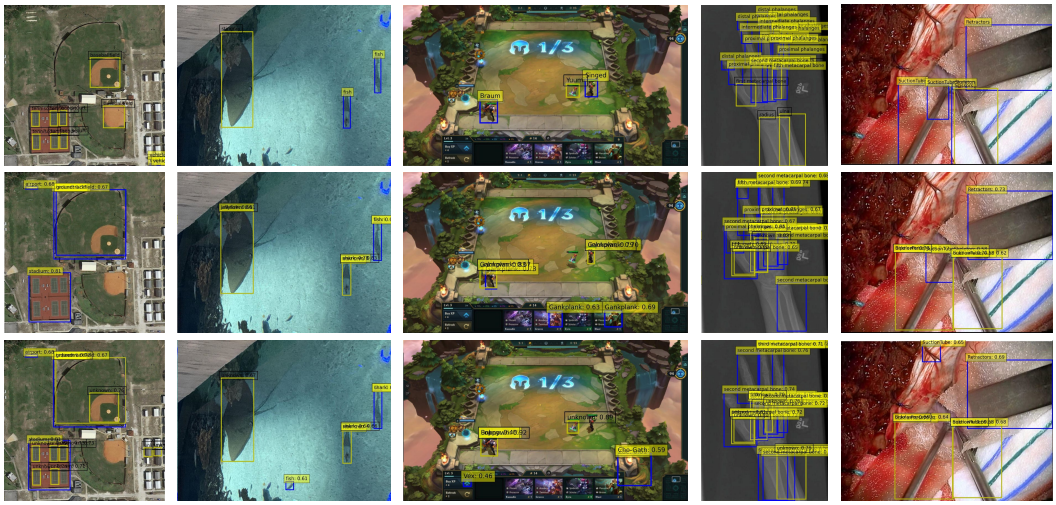

Figure 5: **Visualization of failure cases and limitations across the five RWD datasets.** From left to right: Aerial, Aquatic, Game, Medical, and Surgery. **Top row:** Ground Truth annotations. **Middle row:** Predictions by the baseline (PASS). **Bottom row:** Predictions by our proposed MGME framework. Blue boxes denote known categories, and yellow boxes denote unknown categories.

E Social Impacts and Limitations

E.1 Social Impacts

The proposed MGME framework significantly enhances the adaptability of object detection systems in unconstrained environments, providing substantial societal benefits across safety-critical domains. By enabling the robust discovery of novel objects—ranging from unidentified marine species in **Aquatic** conservation to elusive anatomical structures in **Medical** and **Surgery** imaging—this work empowers specialized AI assistants to handle out-of-distribution (OOD) scenarios with higher reliability. Furthermore, by strictly preventing manifold collisions through topological pruning, our method reduces the risk of misclassification in autonomous systems (e.g., **Aerial** drone monitoring), thereby strengthening the safety and trustworthiness of AI-driven decision-making in high-stakes real-world applications.

E.2 Limitations and Future Work

Despite the empirical effectiveness and geometric rigor of the MGME framework, several limitations remain to be addressed in future research:

1. **Theoretical Foundations of Geodesic Expansion:** While our multi-prototype geodesic escape is supported by extensive empirical results and Riemannian geometric intuition, we have not provided a formal theoretical proof of the convergence of generalization risk within the residual subspace. Future work will focus on establishing the mathematical foundations.
2. **Scalability to Ultra-Dense Scenes:** Although MGME excels at discovering diverse unknowns, its performance can be constrained in extremely crowded scenes where the residual subspace becomes highly saturated. In such cases, satisfying the hard topological constraint may force the generated prototypes too far from the actual distribution, slightly degrading recall. We plan to investigate more adaptive safety margins that dynamically adjust to local feature density.
3. **Impact of VLM Scaling:** Our study validates the effectiveness of MGME on ViT-B/16 and ViT-L/14 architectures. However, the impact of scaling up to even larger vision-language models (e.g., ViT-G or models with billions of parameters) remains unexplored. Future research is needed to investigate whether larger embedding spaces provide a more structured "null-space" for exploration or if they introduce new forms of anisotropic noise.
4. **Integration with Continuous Domain Adaptation:** Given the ongoing evolution of data distributions in the open world, integrating advanced Domain Adaptation techniques with our manifold-guided approach appears promising. Exploring how MGME can autonomously refine its semantic attractors through continuous interaction with streaming real-world data will be a key direction for our future work.
5. **Dependence on a proprietary LLM.** Semantic steering relies on a closed and nondeterministic LLM. The generated concepts may vary with model versions, decoding parameters, and prompt wording, and API usage introduces additional cost. We therefore report the exact model and prompts and use a fixed, pre-generated concept list for all experiments.