OpenPanoD: Aligning Multimodal Prompts and Spherical Representations for Open-Vocabulary Panoramic Detection

Hang Xu[1,2]

[1] Hangzhou Dianzi University, Hangzhou Zhejiang 310018, China
[2] Zhejiang Provincial Key Laboratory of Low Altitude Ubiquitous Networking Technology, Hangzhou Dianzi University, Hangzhou Zhejiang 310018, China
hxu@hdu.edu.cn

Abstract. Panoramic object detection has made steady progress, but existing methods remain largely closed-set and therefore cannot recognize novel categories in open environments. Extending open-vocabulary detection to panoramas is challenging because prompt-image alignment is weakened by two factors: semantic ambiguity in language-only prompts and severe geometric distortion caused by equirectangular projection. We present OpenPanoD, a framework for open-vocabulary panoramic detection that aligns multimodal prompts with geometry-aware spherical representations. On the prompt side, a multimodal prompt encoder combines language descriptions with visual exemplars to obtain category embeddings that are both generalizable and visually specific. On the image side, GeoFormer maps panoramic features onto a quasi-uniform spherical grid and performs spherical attention to reduce projection-induced distortion and boundary discontinuity. The resulting prompt and image embeddings are fused in a unified detection head for BFoV/RBFoV prediction. Experiments on 360-Indoor and PANDORA show that OpenPanoD consistently improves novel-category performance over strong open-vocabulary and visual-prompt baselines. These results demonstrate the importance of jointly addressing semantic ambiguity and panoramic geometry for open-vocabulary detection in 360-degree scenes.

Keywords: Panoramic · Object detection · Open-Vocabulary

1 Introduction

Object detection in panoramic images is a key component for many applications such as virtual reality [22], autonomous driving [2] and intelligent surveillance [24]. However, current detectors [7,5,32,20,31] are confined to a **closed-set** paradigm, recognizing only predefined categories. This severely limits their utility in the unpredictable real world. The crucial next step is to build **open-vocabulary** detectors, yet this transition is uniquely challenging for panoramic images. Success in open-vocabulary detection hinges on the precise **alignment** between arbitrary prompts (what to find) and image features (where it appears).

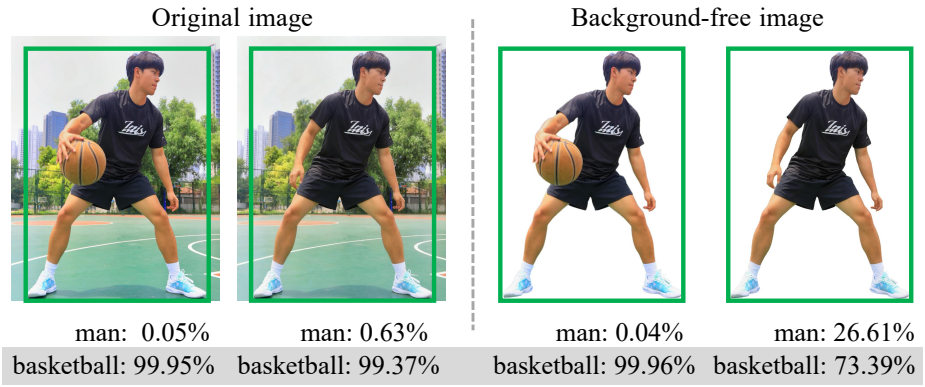

Fig. 1: Language-only CLIP prompts can be biased by contextual co-occurrence. In this example, the prompt *basketball* receives much higher confidence than *person/man*, illustrating semantic ambiguity in prompt-based recognition.

In panoramic imagery, this alignment process is compromised by two intertwined challenges: semantic ambiguity in prompts and geometric distortion of image features.

First, the **semantic ambiguity** of prompts limits detection robustness. Prevailing open-vocabulary methods rely on text prompts processed by vision-language models such as CLIP [26]. While powerful, these models may fail in complex scenes by learning spurious contextual correlations. For instance, as shown in Fig. 1, when a region contains both a person and a basketball, CLIP-based image-text similarity may assign a much higher score to the co-occurring object label *basketball* than to *person/man*. This indicates that text-only prompts can be biased by contextual co-occurrence rather than focusing on the intended target object. This problem is not unique to panoramas, but it becomes more pronounced in panoramic scenes, where dense context and frequent object co-occurrence make language-only prompts more ambiguous. Conversely, vision-prompt-based methods [14], which use image exemplars, offer visual specificity but lack semantic generalization; defining an abstract concept such as *bird* would require an impractical number of examples. This reveals a clear complementarity: text provides generalization, while vision offers specificity.

Second, and more critically for panoramas, **geometric corruption** weakens the visual side of the alignment. Equirectangular projection (ERP) severely distorts objects, altering their shape, scale, and structural integrity. This means even with a reliable prompt, the corresponding image features can be warped and unreliable. The issue is **especially detrimental in an open-vocabulary setting**. For base classes, a model might learn distorted appearances through extensive training. For *novel* classes, however, the model has no prior knowledge of how a distorted novel object should appear, making the alignment between a clean semantic prompt and a warped image feature difficult. Robust alignment therefore requires geometry-aware spherical image features.

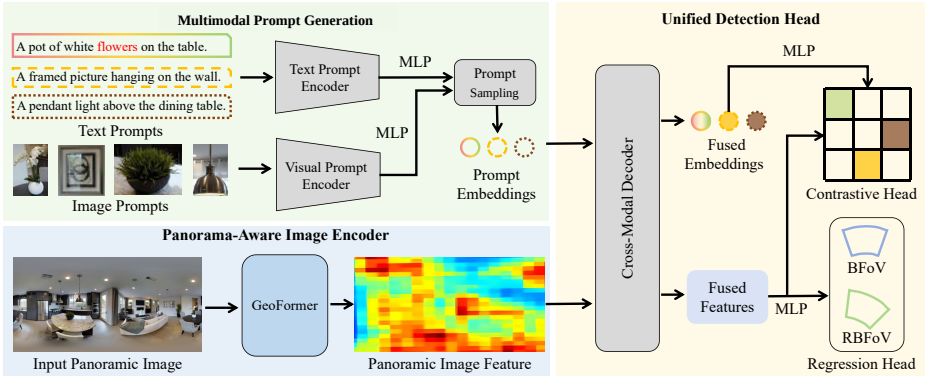

Fig. 2: **The overall architecture of our proposed OpenPanoD.** The model takes a panoramic image and multimodal (text and vision) prompts as input. A panorama-aware encoder (**GeoFormer**) extracts geometry-aware spherical features, while a multimodal prompt encoder generates unified prompt embeddings. These are then fused in the detection head for open-vocabulary panoramic object detection.

To address this dual challenge, we introduce **OpenPanoD**, a unified framework for aligning multimodal prompts with spherical panoramic representations. The key idea is to make both sides of open-vocabulary alignment reliable: multimodal prompts reduce semantic ambiguity on the query side, while GeoFormer produces geometry-aware spherical features on the image side. As illustrated in Fig. 2, OpenPanoD co-designs two core components. First, a **multimodal prompt encoder** fuses text and visual prompts to generate category embeddings that are both generalizable and visually specific. Second, a panorama-aware image encoder, **GeoFormer**, uses a Transformer encoder on a quasi-uniform spherical grid to reduce projection-induced distortion and boundary discontinuity. These two components work in concert, ensuring that rich semantic guidance is matched against geometry-aware image features for precise open-vocabulary detection.

In this work, the detector is trained only with annotations from base categories. At inference time, it receives text prompts, visual prompts, or their combination for both base and unseen novel categories, and predicts BFoV/RBFoV boxes under the same spherical IoU evaluation protocol.

Evaluated in this zero-shot setting on panoramic detection benchmarks, OpenPanoD achieves state-of-the-art performance. Our contributions are:

- We formulate open-vocabulary panoramic detection as a prompt-feature alignment problem under both semantic ambiguity and spherical geometric distortion, and propose **OpenPanoD** to jointly address the two factors.
- We introduce a **multimodal prompt encoder** that combines category-level language descriptions with visual exemplars, yielding prompt embeddings that are both generalizable and visually specific.

- We introduce **GeoFormer**, a panorama-aware image encoder built on a quasi-uniform icosahedral grid, and validate that geometry-aware spherical representations improve novel-category detection on BFoV and RBFoV benchmarks.
- We conduct extensive experiments that not only establish a new state-of-the-art but also systematically validate the necessity of both multimodal prompts and spherical representations for this challenging task.

2 Related Work

2.1 Panoramic Object Detection

Object detection in 360° images presents unique challenges compared to planar images, primarily due to severe geometric distortions near the poles, object wrap-around at image boundaries, and extreme variations in scale. Existing approaches to tackle these issues can be broadly categorized as follows:

Planar-based Adaptation. Early works focused on adapting standard 2D detectors like Faster R-CNN [28] and YOLO [27] to panoramic data. This was typically achieved by applying them to equirectangular projections (ERP) [5,32,31] or other projection formats like tangent planes [33,21]. However, these methods force planar models to contend with heavily distorted image features, fundamentally limiting their representation quality and robustness.

Specialized Spherical Encoders. A more principled approach involves designing encoders that operate directly on the spherical manifold. This includes Spherical CNNs [6,13], which define convolutions on spherical grids. While an improvement, they often introduce approximation errors and their fixed, local receptive fields fail to model complex geometric structures effectively. More recently, Spherical Transformers [18,1] have emerged, but they typically resort to local attention mechanisms to manage computational cost. This compromises the Transformer's core strength of direct long-range modeling, as global context is only aggregated indirectly.

Despite these architectural advances, all prior work is confined to a closed-set paradigm, rendering them unable to detect novel objects and limiting their practical use.

2.2 Open-Vocabulary Object Detection

Open-vocabulary object detection (OVD) aims to overcome the closed-set limitation by enabling models to detect objects from an arbitrary, open-ended set of categories defined by natural language text. The field gained significant momentum with the advent of powerful vision-language pre-trained models like CLIP [26]. Current OVD methods can be broadly categorized into two main architectural approaches:

Two-Stage Alignment Architectures. These methods, such as ViLD [8] and OWL-ViT [25], first use a standard detector to propose regions and then

employ a CLIP-like model to classify these regions based on text prompts. They typically align features at a late stage, offering efficiency but sometimes sacrificing fine-grained localization accuracy.

End-to-End Fusion Architectures. This dominant approach, exemplified by GLIP [15], GroundingDINO [19], and YOLO-World [4], performs deep, early fusion between image and prompt features using cross-attention mechanisms. This allows the text prompt to guide the feature extraction and localization process, leading to superior performance in fine-grained grounding.

While powerful, these methods rely exclusively on textual prompts to define novel categories.

2.3 Multimodal Prompting for Object Detection

Recognizing the limitations of text-only prompts, recent work [14] uses one or more support images to define and locate a target category. The natural evolution is multimodal prompting, which seeks to combine the general semantic guidance of text with the specific appearance details from visual exemplars.

Several prominent works, including T-Rex2 [12], DINO-X [29], and Prompt-DINO [9], explore text-visual prompting and demonstrate its potential for open-world detection and segmentation. However, these systems are primarily designed for standard planar images, where the visual features are not affected by spherical projection distortion. Directly applying their prompting strategy to panoramas is therefore suboptimal: even informative prompts can be poorly aligned with ERP-distorted image features. In contrast, OpenPanoD studies multimodal prompting under panoramic geometry and couples prompt fusion with spherical feature learning. This co-design addresses both open-vocabulary generalization and panoramic distortion.

3 Methodology

3.1 Overview

We introduce OpenPanoD, an end-to-end framework for open-vocabulary object detection in panoramic images. During training, prompts are constructed only for base categories and used as category-conditioned queries and classification prototypes. The CLIP text and image encoders are kept frozen, and only the prompt projection layers, GeoFormer, cross-modal decoder, and detection heads are optimized with base-category annotations. At inference time, text prompts, visual prompts, or their combination are used to specify both base and unseen novel categories. As illustrated in Fig. 2, our model consists of three core components:

– **A Multimodal Prompting Mechanism** that generates rich text and vision embeddings to represent object categories, providing a flexible and extensible vocabulary for the detector.

- **A Panorama-Aware Image Encoder** that operates on a quasi-uniform spherical grid derived from a subdivided icosahedron. This allows the model to process image features in a distortion-aware manner, preserving spatial relationships across the entire sphere.
- **A Unified Detection Head** that fuses image features with prompt embeddings via cross-attention. It is trained with a combination of classification, regression, and supervised contrastive losses to align image features with the multimodal semantic space, enabling robust open-vocabulary performance.

3.2 Multimodal Prompt Generation

To equip our model with rich open-vocabulary understanding and fine-grained visual recognition capabilities, we devise a multimodal prompting mechanism that leverages both text and vision prompts. The generation and encoding process for these prompts is performed offline to ensure training efficiency.

Text Prompt Representation. For each object category, we generate a set of semantically diverse text descriptions to create a robust conceptual representation. Specifically, we employ the DeepSeek [10] language model to generate a candidate pool of descriptive phrases and use $K_t = 10$ text prompts by default, following the prompt-quantity analysis in Sec. 4. These phrases range from simple category names (e.g., "a chair") to more elaborate descriptions (e.g., "a piece of furniture for sitting" and "a seat with a back and four legs").

Each phrase is then encoded into an embedding using the pre-trained CLIP [26] text encoder. To form a single, robust text representation e_t for the category, we compute the element-wise average of the $N = K_t$ phrase embeddings:

$$e_t = \frac{1}{N} \sum_{i=1}^{N} \text{CLIP}_{\text{text}}(\text{phrase}_i) \tag{1}$$

where N is the number of generated phrases for the category. This averaging process distills a generalized semantic embedding that is less sensitive to the peculiarities of any single description.

Visual Prompt Representation. To complement the abstract nature of text prompts, we construct a visual prompt library that provides concrete visual exemplars for each category. The visual prompt library is constructed without using any panoramic training or test images, ensuring that no images from the panoramic detection benchmarks are introduced into the prompt library. For each category, we first map its category name to ImageNet-21K synsets using exact name matching and WordNet aliases, and retrieve up to 100 candidate images from the matched synsets. We then rank the retrieved candidates according to the available category supervision. If the category has a corresponding ImageNet-1K classifier label, we rank the candidate images using the classification logit of a standard ResNet-50 [11] classifier pre-trained on ImageNet-1K.

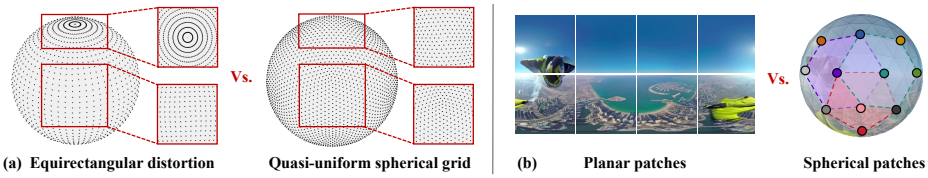

Fig. 3: Illustration of (a) different sampling strategies on the spherical surface and (b) comparison between planar patches and spherical patches.

For categories without a direct ImageNet-1K classifier label, we instead rank the candidates by CLIP image-text similarity between each candidate image and the category text prompt. Finally, we keep the top $K_v = 5$ images by default, which provides the best trade-off in Sec. 4.

These selected images are then processed by the pre-trained CLIP image encoder to obtain their respective embeddings. Similar to text prompt generation, we average the $M = K_v$ image embeddings to create a single canonical visual prompt embedding e_v:

$$e_v = \frac{1}{M} \sum_{j=1}^{M} \text{CLIP}_{\text{image}}(\text{image}_j), \tag{2}$$

where $M = K_v$. The resulting embedding e_v captures the prototypical visual characteristics of the category and provides visually grounded guidance complementary to the semantic information from text prompts.

Unified Prompt Projection. The text and vision embeddings, e_t and e_v, reside in the original CLIP feature space. Although this space provides an initial semantic alignment, it is primarily optimized for image-level representation learning rather than fine-grained object localization. To adapt these general-purpose embeddings to open-vocabulary detection, we introduce a projection stage that maps them into a task-specific prompt space.

Specifically, we use two separate Multi-Layer Perceptron (MLP) networks, one for each modality. These MLPs serve as learnable adapters that project the high-dimensional CLIP embeddings into a unified 256-dimensional detection-oriented space. Each MLP consists of two fully connected layers with a ReLU activation function in between. For category c, the projection is formulated as:

$$p_c^t = \text{MLP}_{\text{text}}(e_c^t), \quad p_c^v = \text{MLP}_{\text{vision}}(e_c^v), \tag{3}$$

where $p_c^t, p_c^v \in \mathbb{R}^{256}$ denote the projected text and visual prompt embeddings, respectively. Using separate MLPs allows the model to learn modality-specific transformations while mapping both modalities into a shared embedding space optimized for detection.

To explicitly define how different prompt modalities are used, we construct a prompt set for each category:

$$\mathcal{P}_c = \begin{cases} \{p_c^t\}, & \text{text-only setting,} \\ \{p_c^v\}, & \text{visual-only setting,} \\ \{p_c^t, p_c^v\}, & \text{multimodal setting.} \end{cases} \tag{4}$$

For the decoder query, we use the averaged category prompt:

$$\bar{p}_c = \frac{1}{|\mathcal{P}_c|} \sum_{p \in \mathcal{P}_c} p. \tag{5}$$

During classification, however, we retain all prompt variants in $\mathcal{P}_c$ and compute the category score by taking the maximum similarity over these prompts. This design allows the decoder to use a compact category-level query while preserving modality-specific prompt cues for classification.

3.3 Panorama-Aware Image Encoder

To effectively process 360° images while alleviating the distortion and boundary artifacts introduced by equirectangular projection (ERP), we design a panorama-aware image encoder, termed GeoFormer, as illustrated in Fig. 4. Instead of applying attention directly to uniformly sampled ERP patches, GeoFormer first maps panoramic features onto a quasi-uniform spherical grid and then performs feature interaction over spherical tokens. This design reduces the bias caused by non-uniform ERP sampling and provides geometry-aware representations for prompt-image alignment. The encoder consists of two main stages: spherical grid construction and stacked spherical attention layers.

Spherical Grid Construction. We discretize the sphere using a subdivided icosahedron, which provides a quasi-uniform set of vertices on the spherical surface, as shown in Fig. 3(a). Starting from a base icosahedron with 20 faces and 12 vertices, we iteratively subdivide its triangular faces and normalize the generated vertices onto the unit sphere. Compared with uniform sampling in the ERP plane, this icosahedral discretization produces spherical patches with more balanced solid angles, thereby reducing the shape and area distortion caused by ERP, as illustrated in Fig. 3(b).

To initialize spherical tokens, we transfer features from the ERP feature map to the icosahedral grid. For each spherical vertex $\mathbf{s}_i = (x_i, y_i, z_i)$, we compute its longitude and latitude and then project it to ERP coordinates:

$$\theta_i = \operatorname{atan2}(y_i, x_i), \quad \phi_i = \arcsin(z_i), \tag{6}$$

$$u_i = \frac{\theta_i + \pi}{2\pi} W, \quad \nu_i = \left(\frac{1}{2} - \frac{\phi_i}{\pi}\right) H, \tag{7}$$

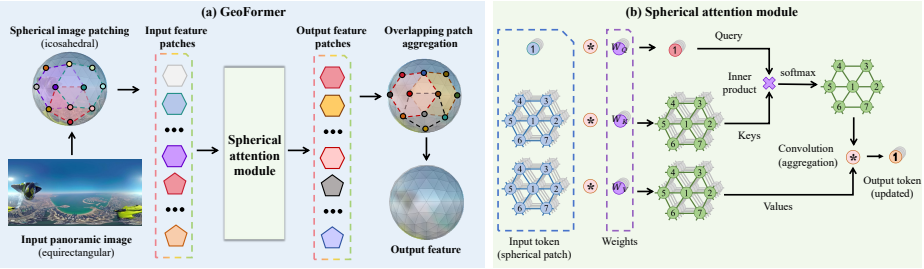

Fig. 4: Illustration of GeoFormer. (a) GeoFormer maps ERP features onto a quasi-uniform spherical grid and performs feature interaction over spherical tokens. (b) The spherical attention layer enables global information exchange among spherical tokens.

where W and H denote the width and height of the ERP feature map, respectively. Here, u_i and ν_i denote the horizontal and vertical ERP coordinates corresponding to the spherical vertex $\mathbf{s}_i$. We then use bilinear sampling with horizontal circular padding to obtain the initial spherical token feature $\mathbf{x}_i^0$ from the ERP feature map at location (u_i, ν_i). The horizontal circular padding preserves the wrap-around continuity of panoramic images.

To encode the spatial layout of spherical tokens, we further inject 3D positional information. Specifically, the Cartesian coordinates of each spherical vertex are projected by a small MLP and added to the sampled feature:

$$\mathbf{x}_i^{\text{in}} = \mathbf{x}_i^0 + \text{MLP}_{\text{pos}}([x_i, y_i, z_i]), \tag{8}$$

where MLP_{pos} maps the 3D coordinate to the same feature dimension as $\mathbf{x}_i^0$. The resulting token set serves as the input to the spherical attention layers.

Spherical Attention Layer. Given the initialized spherical tokens, GeoFormer applies a stack of L spherical attention layers. Each layer follows the Transformer encoder design but operates on tokens sampled from the spherical grid rather than regular ERP patches. This allows information to be exchanged globally across the panoramic scene while preserving the spherical geometry encoded by the sampling grid and 3D positional embeddings.

Let $\mathbf{X}^{l-1} \in \mathbb{R}^{N_v \times d}$ denote the input token features of the l-th layer, where N_v is the number of spherical vertices and d is the feature dimension. The query, key, and value matrices are computed as:

$$\mathbf{Q} = \mathbf{X}^{l-1}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}^{l-1}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}^{l-1}\mathbf{W}_V, \tag{9}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V \in \mathbb{R}^{d \times d}$ are learnable projection matrices.

For each attention head, the scaled dot-product attention is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{10}$$

where $d_k = d/h$ is the channel dimension of each head and h is the number of attention heads. The outputs of all heads are concatenated and linearly projected:

$$\mathrm{MSA}(\mathbf{X}^{l-1}) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)\mathbf{W}_O, \tag{11}$$

where $\mathbf{W}_O \in \mathbb{R}^{d\times d}$ is the output projection matrix.

Following the standard Transformer architecture, each spherical attention layer uses residual connections, layer normalization, and a feed-forward network:

$$\widetilde{\mathbf{X}}^l = \mathbf{X}^{l-1} + \mathrm{MSA}(\mathrm{LN}(\mathbf{X}^{l-1})), \tag{12}$$

$$\mathbf{X}^l = \widetilde{\mathbf{X}}^l + \mathrm{MLP}(\mathrm{LN}(\widetilde{\mathbf{X}}^l)). \tag{13}$$

By stacking these layers, GeoFormer progressively builds geometry-aware panoramic features, which are better aligned with multimodal category prompts in the subsequent detection head.

3.4 Unified Detection Head

The object detection head takes the panoramic features from GeoFormer and the multimodal prompt embeddings as input, and fuses them to produce final bounding box predictions and class labels. It consists of a prompt-conditioned cross-modal decoder, decoupled prediction heads, and a multi-component loss function.

Cross-Modal Decoder. We employ a Transformer-based decoder to fuse image and prompt information. Let $\mathbf{X}_{\mathrm{img}} \in \mathbb{R}^{N_p\times d}$ denote the final panoramic feature map from GeoFormer, where N_p is the number of image patches. Let $\{\bar{\mathbf{p}}_c\}_{c=1}^{N_c}$ denote the averaged category prompts defined in Eq. 5, where N_c is the number of prompted categories.

To support multiple instances of the same category, the prompt-sampling module expands each averaged category prompt into K prompt-conditioned object queries by adding learnable instance offsets:

$$\mathbf{P}^0_{c,k} = \bar{\mathbf{p}}_c + \mathbf{o}_k, \quad c = 1, \ldots, N_c,\ k = 1, \ldots, K, \tag{14}$$

where $\bar{\mathbf{p}}_c$ is the averaged category prompt and $\mathbf{o}_k$ is a learnable instance offset. The CLIP prompt embeddings are generated offline, while the projection MLPs and instance offsets are optimized during detector training. This design preserves category-specific prompt guidance while allowing multiple detections for each category.

The decoder is composed of a stack of L identical layers. Each layer first lets the prompt-conditioned object queries attend to panoramic image features to gather visual evidence, and then lets the queries interact with each other

to refine their contextual representations. A single decoder layer l operates as follows:

$$\mathbf{P}' = \mathrm{LN}\left(\mathbf{P}^{l-1} + \mathrm{MHCA}(\mathbf{P}^{l-1}, \mathbf{X}_{\mathrm{img}}, \mathbf{X}_{\mathrm{img}})\right), \tag{15}$$

$$\mathbf{P}'' = \mathrm{LN}\left(\mathbf{P}' + \mathrm{MHSA}(\mathbf{P}', \mathbf{P}', \mathbf{P}')\right), \tag{16}$$

$$\mathbf{P}^{l} = \mathrm{LN}\left(\mathbf{P}'' + \mathrm{MLP}(\mathbf{P}'')\right), \tag{17}$$

where $\mathbf{P}^{l-1}$ denotes the object queries from the previous layer, MHCA denotes Multi-Head Cross-Attention, MHSA denotes Multi-Head Self-Attention, and LN denotes Layer Normalization. The output of the final decoder layer, $\mathbf{P}^{L}$, contains context-rich representations for each prompt-conditioned object query.

Prediction Heads. The output queries $\mathbf{P}^{L}$ are fed into two decoupled prediction heads that share weights across all queries:

Regression Head: A 3-layer MLP with ReLU activations predicts the BFoV/RBFoV coordinates for each query.

Classification Head: A linear layer projects each output query $\mathbf{P}_i^{L}$ into a 256-dimensional embedding space, producing the final detection embedding $\mathbf{z}_i \in \mathbb{R}^{256}$.

Loss Functions. Our final training objective is a weighted sum of three distinct loss components: a classification loss, a supervised contrastive loss, and a bounding box regression loss.

Classification Loss. The classification score $s_{i,c}$ for the i-th prediction and the c-th category is determined by the similarity between its detection embedding $\mathbf{z}_i$ and the corresponding prompt embeddings. We use a scaled and shifted cosine similarity, taking the maximum score over all prompt variants for a given category:

$$s_{i,c} = \max_{j:\mathrm{label}(p_j)=c} \left(\alpha \frac{\mathbf{z}_i \cdot \mathbf{p}_j}{\|\mathbf{z}_i\| \|\mathbf{p}_j\|} + \beta \right) \tag{18}$$

where $\mathbf{p}_j$ is a prompt embedding (either text or vision), and α, β are learnable parameters that scale and shift the logits. We then apply a standard Focal Loss [16] to these scores to handle class imbalance, denoted as $\mathcal{L}_{\mathrm{cls}}$.

Supervised Contrastive Loss. To further enhance the discriminative power of the embedding space, we introduce a supervised contrastive loss, $\mathcal{L}_{\mathrm{ct}}$. This loss encourages each detection embedding $\mathbf{z}_i$ to be closer to its corresponding positive class prompts than to any negative prompts. For a given prediction $\mathbf{z}_i$ matched to a ground-truth object of class c, its positive set $\mathcal{P}_i$ consists of all prompt embeddings $\{\mathbf{p}_j\}$ belonging to class c. The negative set $\mathcal{N}_i$ contains all prompts from other classes. The loss is formulated as:

$$\mathcal{L}_{\mathrm{ct}} = \sum_{i \in \mathcal{M}} -\log \frac{\sum_{\mathbf{p}^+ \in \mathcal{P}_i} e^{\mathrm{sim}(\mathbf{z}_i, \mathbf{p}^+)/\tau}}{\sum_{\mathbf{p}^+ \in \mathcal{P}_i} e^{\mathrm{sim}(\mathbf{z}_i, \mathbf{p}^+)/\tau} + \sum_{\mathbf{p}^- \in \mathcal{N}_i} e^{\mathrm{sim}(\mathbf{z}_i, \mathbf{p}^-)/\tau}} \tag{19}$$

where $\mathcal{M}$ is the set of matched predictions, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, and τ is a temperature hyperparameter (set to 0.07).

Bounding Box Loss. The box loss $\mathcal{L}_{\text{box}}$ is a linear combination of the L1 loss and the Gaussian Label Distribution Learning (GLDL) loss [31], applied to the predictions matched with ground-truth boxes:

$$\mathcal{L}_{\text{box}} = \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} + \lambda_{\text{GLDL}}\mathcal{L}_{\text{GLDL}} \tag{20}$$

Total Loss. The final loss function is a weighted sum of the three components:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{ct}}\mathcal{L}_{\text{ct}} + \lambda_{\text{box}}\mathcal{L}_{\text{box}} \tag{21}$$

where λ_{cls}, λ_{ct}, and λ_{box} are hyper-parameters that balance the contribution of each loss term.

4 Experiments

4.1 Experimental Setup

Datasets and Preprocessing. We conduct experiments on two panoramic detection benchmarks: **360-Indoor** [5] contains 3,335 real-world indoor images with 89,148 bounding field of view (BFoV) annotations across 37 categories. We designate 30 as base and 7 as novel categories. **PANDORA** [32] consists of 3,000 panoramic images from real-world indoor scenarios with 94,353 rotated bounding field of view (RBFoV) annotations across 47 categories. We use 40 base and 7 novel categories.

We follow the common 1/2, 1/6, and 1/3 train/validation/test split protocol and use the same split for all methods. All images are resized to 512×1024 for both training and inference. During training, only annotations from base classes are used.

Evaluation Metrics. We report mean Average Precision (mAP) [17] at a Sphere-IoU [7] threshold of 0.5. Performance is measured on base categories (mAP_{base}), novel categories ($\text{mAP}_{\text{novel}}$), and their Harmonic Mean (HM), which provides a balanced view of overall performance.

Implementation Details. Our framework is implemented using MMDetection 3.3.0 [3]. Unless otherwise specified, we use $K_t = 10$ text prompts and $K_v = 5$ visual prompts per category. We use the AdamW [23] optimizer with an initial learning rate of 2.5×10^{-4} and a weight decay of 0.05. A cosine annealing scheduler is employed. All experiments are run on 8 NVIDIA 3090Ti GPUs with a total batch size of 16.

Baselines. We compare against seven text prompt-based detectors: ViLD [8], OV-DETR [34], GroundingDINO [19], GLIP [15], BARON [30], YOLO-World [4], and Prompt-DINO [9], and two visual prompt-based detectors: DINOv [14] and T-Rex2 [12]. Since these models were designed for planar images, we adapt them by modifying their prediction heads to support BFoV/RBFoV regression and fine-tune them on the same base-class annotations, image resolution, and

Table 1: Comparison with the state-of-the-art detectors.

Dataset	Method	Prompt Type	Backbone	mAP_{base}	mAP_{novel}	HM
360-Indoor [5]	ViLD [8]	Text Prompt	R50	17.4	13.8	15.4
	OV-DETR [34]	Text Prompt	R50	21.1	19.2	20.1
	GroundingDINO [19]	Text Prompt	R50	26.5	20.8	23.3
	GLIP [15]	Text Prompt	Swin-T	27.2	20.5	23.4
	BARON [30]	Text Prompt	R50	25.4	19.1	21.8
	YOLO-World [4]	Text Prompt	YOLOv8-M	27.9	26.6	27.2
	Prompt-DINO [9]	Text Prompt	ViT-L	28.3	26.8	27.5
	OpenPanoD	Text Prompt	GeoFormer	**30.5**	**29.8**	**30.1**
	DINOv [14]	Visual Prompt	Swin-L	22.3	20.7	21.5
	T-Rex2 [12]	Visual Prompt	Swin-L	21.4	19.6	20.5
	OpenPanoD	Visual Prompt	GeoFormer	**25.9**	**25.5**	**25.7**
	OpenPanoD	Multimodal Prompt	GeoFormer	**35.3**	**33.9**	**34.6**
PANDORA [32]	ViLD [8]	Text Prompt	R50	14.5	11.1	12.6
	OV-DETR [34]	Text Prompt	R50	19.6	17.3	18.4
	GroundingDINO [19]	Text Prompt	R50	23.9	18.8	21.0
	GLIP [15]	Text Prompt	Swin-T	23.4	19.2	21.1
	BARON [30]	Text Prompt	R50	23.7	18.7	20.9
	YOLO-World [4]	Text Prompt	YOLOv8-M	26.8	25.3	26.0
	Prompt-DINO [9]	Text Prompt	ViT-L	27.1	25.5	26.3
	OpenPanoD	Text Prompt	GeoFormer	**29.6**	**28.7**	**29.1**
	DINOv [14]	Visual Prompt	Swin-L	20.8	18.9	19.8
	OpenPanoD	Visual Prompt	GeoFormer	**25.4**	**24.9**	**25.1**
	OpenPanoD	Multimodal Prompt	GeoFormer	**34.4**	**31.3**	**32.8**

train/validation/test split. For T-Rex2, the official API returns horizontal boxes; therefore, we report it on the BFoV-based 360-Indoor benchmark and omit it from PANDORA's rotated RBFoV setting.

4.2 Comparison with State-of-the-Art Methods

Tab. 1 compares OpenPanoD with text-prompt and visual-prompt open-vocabulary detectors adapted to panoramic detection. On **360-Indoor**, the multimodal version of OpenPanoD achieves 33.9% mAP_{novel} and 34.6 HM, outperforming the strongest baseline, Prompt-DINO, by 7.1 points in novel mAP. On **PANDORA**, where objects require rotated bounding-field-of-view prediction, OpenPanoD achieves 31.3% mAP_{novel} and 32.8 HM, improving over Prompt-DINO by 5.8 points. The gains are consistent across text-only, visual-only, and multimodal settings, suggesting that the proposed spherical representation benefits prompt-based detection beyond a specific prompt type.

4.3 Ablation Studies

Component-wise Analysis. As shown in Tab. 2, we isolate the effect of each component on PANDORA. The first three rows use a planar ERP encoder, while the final row replaces it with GeoFormer. Starting from a text-prompt baseline, adding visual prompts improves mAP_{novel} from 24.9% to 26.1%, showing the

Table 2: Ablation study of various components on the PANDORA dataset. Rows without the spherical encoder use a planar ERP encoder.

Text Prompts	Visual Prompts	Contrastive Loss	Spherical Encoder	mAP_{base}	mAP_{novel}	HM
✓				27.6	24.9	26.2
✓	✓			29.7	26.1	27.8
✓	✓	✓		31.5	29.6	30.5
✓	✓	✓	✓	34.4	31.3	32.8

Table 3: Ablation study of image encoders.

Image Encoder	mAP_{base}	mAP_{novel}
ViT	31.5	29.6
SphereUFormer	33.1	30.2
GeoFormer	34.4	31.3

Table 4: Ablation study of prompt quantity.

Prompt Type	#Number of Prompts 1	3	5	10	20
360-Indoor					
Image Prompt	24.3	24.8	**25.5**	25.2	24.7
Text Prompt	24.6	25.2	25.9	**26.4**	26.3
PANDORA					
Image Prompt	22.1	22.9	**23.3**	22.7	22.3
Text Prompt	23.4	23.8	24.5	**24.9**	24.3

complementary value of visual exemplars. Adding the supervised contrastive loss further improves mAP_{novel} to 29.6% by explicitly separating matched prompt-object pairs from negative prompts. Finally, replacing the planar encoder with GeoFormer raises mAP_{novel} to 31.3%, confirming that geometry-aware spherical features are critical for open-vocabulary panoramic detection.

Analysis of Image Encoder Architectures. We compare different image encoder architectures: A standard Vision Transformer (ViT) applied directly to the ERP image; A competing spherical encoder, SphereUFormer [1]; and our GeoFormer. As shown in Tab. 3, our method outperforms both the ViT (+1.7% mAP_{novel}) and SphereUFormer (+1.1% mAP_{novel}). This suggests that our GeoFormer provides a more effective and geometrically sound way to capture features across the entire sphere compared to other methods, leading to more robust representations.

Impact of Prompt Quantity. Table 4 analyzes the sensitivity to the number of text and vision prompts. For text prompts, performance saturates around 10 prompts, indicating that sufficient semantic context is beneficial. For vision prompts, we observe peak performance with 5 samples; more samples lead to a slight degradation, likely due to the introduction of noisy or redundant visual information. These findings offer practical guidelines for prompt engineering in vision-language models for detection.

4.4 Qualitative Analysis

Fig. 5 visualizes the detection results of OpenPanoD. The visualizations demonstrate our model's ability to accurately localize and classify both base and novel

Fig. 5: Qualitative detection results. Top: horizontal BFoV detection results. Bottom: oriented RBFoV detection results.

objects, even those undergoing severe distortion near the poles or wrapping around the image boundary.

5 Conclusion

In this work, we introduced OpenPanoD, a framework for open-vocabulary object detection in panoramic images. We identified that this task is hindered by a dual challenge: semantic ambiguity on the prompt side and geometric corruption on the image side. OpenPanoD addresses these issues by combining a multimodal prompt encoder with GeoFormer, a panorama-aware image encoder that produces geometry-aware spherical features for robust prompt-feature alignment. Extensive experiments demonstrate that OpenPanoD establishes a new state-of-the-art and significantly outperforms prior methods on novel category detection. While our work sets a new benchmark, future research could explore optimizing its computational efficiency or extending the framework to broader panoramic understanding tasks, such as segmentation, grounding, and video understanding. We believe OpenPanoD provides an effective framework for this task and lays a robust foundation for future research into open-vocabulary perception in 360-degree environments.

Acknowledgements

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LQN25F020015. The authors acknowledge the Supercomputing Center of Hangzhou Dianzi University for providing computing resources.

References

1. Benny, Y., Wolf, L.: SphereUFormer: A u-shaped transformer for spherical 360 perception. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 940–950 (2025)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027 (2019)
3. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
4. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: YOLO-World: Real-time open-vocabulary object detection. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2024)
5. Chou, S.H., Sun, C., Chang, W.Y., Hsu, W.T., Sun, M., Fu, J.: 360-Indoor: Towards learning real-world objects in 360 indoor equirectangular images. In: IEEE Winter Conference on Applications of Computer Vision. pp. 834–842 (2020)
6. Coors, B., Condurache, A.P., Geiger, A.: SphereNet: Learning spherical representations for detection and classification in omnidirectional images. In: Proceedings of the European conference on computer vision (ECCV). pp. 518–533 (2018)
7. Dai, F., Chen, B., Xu, H., Ma, Y., Li, X., Feng, B., Yan, C., Zhao, Q.: Unbiased IoU for spherical image object detection. In: AAAI (2022)
8. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)
9. Guan, Y., Sun, C., Fu, C., Huang, Z., Yuan, C., Li, C.: Text-guided visual prompt DINO for generic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21288–21298 (2025)
10. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
12. Jiang, Q., Li, F., Zeng, Z., Ren, T., Liu, S., Zhang, L.: T-Rex2: Towards generic object detection via text-visual prompt synergy. In: European Conference on Computer Vision. pp. 38–57. Springer (2024)
13. Lee, Y., Jeong, J., Yun, J., Cho, W., Yoon, K.J.: SpherePHD: Applying CNNs on a spherical polyhedron representation of 360 images. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 9173–9181 (2019)
14. Li, F., Jiang, Q., Zhang, H., Ren, T., Liu, S., Zou, X., Xu, H., Li, H., Yang, J., Li, C., et al.: Visual in-context prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12861–12871 (2024)
15. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10955–10965 (2022). https://doi.org/10.1109/CVPR52688.2022.01069

16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988 (2017)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. pp. 740–755. Springer (2014)
18. Ling, Z., Xing, Z., Zhou, X., Cao, M., Zhou, G.: PanoSwin: A pano-style swin transformer for panorama understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17755–17764 (2023)
19. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
20. Liu, X., Xu, H., Chen, B., Zhao, Q., Ma, Y., Yan, C., Dai, F.: Sph2pob: Boosting object detection on spherical images with planar oriented boxes methods. In: IJCAI. pp. 1231–1239 (2023)
21. Liu, X., Xu, H., Chen, B., Zhao, Q., Ma, Y., Yan, C., Dai, F.: Sph2Pob: Boosting object detection on spherical images with planar oriented boxes methods. In: IJCAI. pp. 1231–1239 (2023)
22. Lo, W.C., Fan, C.L., Lee, J., Huang, C.Y., Chen, K.T., Hsu, C.H.: 360° video viewing dataset in head-mounted virtual reality. In: Proc. ACM Conf. on Multimedia Syst. (2017)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
24. Masuyama, Y., Bando, Y., Yatabe, K., Sasaki, Y., Onishi, M., Oikawa, Y.: Self-Supervised Neural Audio-Visual Sound Source Localization via Probabilistic Spatial Modeling. In: IROS (2020)
25. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems **36** (2024)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
27. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788 (2016)
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2017)
29. Ren, T., Chen, Y., Jiang, Q., Zeng, Z., Xiong, Y., Liu, W., Ma, Z., Shen, J., Gao, Y., Jiang, X., et al.: DINO-X: A unified vision model for open-world object detection and understanding. arXiv preprint arXiv:2411.14347 (2024)
30. Wu, S., Zhang, W., Jin, S., Liu, W., Loy, C.C.: Aligning bag of regions for open-vocabulary object detection. In: CVPR (2023)
31. Xu, H., Liu, X., Zhao, Q., Ma, Y., Yan, C., Dai, F.: Gaussian label distribution learning for spherical image object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1033–1042 (June 2023)
32. Xu, H., Zhao, Q., Ma, Y., Li, X., Yuan, P., Feng, B., Yan, C., Dai, F.: PANDORA: A panoramic detection dataset for object with orientation. In: ECCV (2022)

33. Yang, W., Qian, Y., Kämäräinen, J.K., Cricri, F., Fan, L.: Object detection in equirectangular panorama. In: International Conference on Pattern Recognition. pp. 2190–2195 (2018)
34. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary DETR with conditional matching. In: European Conference on Computer Vision. pp. 106–122. Springer (2022)