arXiv:2512.02906v3 [cs.CV] 19 Mar 2026

MRD: Multi-resolution Retrieval-Detection Fusion for High-Resolution Image Understanding

Fan Yang[1], Xingping Dong[2], Xin Yu[3], Wenhan Luo[4], Wei Liu[5], Kaihao Zhang*
[1]HITSZ [2]Wuhan University [3]The University of Queensland [4] HKUST [5] Video Rebirth

Abstract

*Understanding high-resolution (HR) images remains a critical challenge for multimodal large language models (MLLMs). Recent approaches leverage vision-based retrieval-augmented generation (RAG) to retrieve query-relevant crops from HR images, improving understanding capacity of MLLMs. However, this paradigm often leads to object fragmentation, resulting in semantic bias and incomplete retrieval, while also introducing false positives from irrelevant background patches. To address these issues, we propose **Multi-resolution Retrieval-Detection (MRD)**, a training-free framework that enhances HR image understanding from both local and global perspectives. Locally, MRD enforces cross-scale semantic consistency via multi-resolution semantic fusion to mitigate single-resolution bias and alleviate object fragmentation. Globally, it integrates open-vocabulary object detection (OVD) as localization priors within a unified framework. Extensive experiments across multiple MLLMs on HR image benchmarks demonstrate that MRD achieves state-of-the-art (SOTA) performance on both single-object and multi-object understanding tasks. Code will be available at: https://github.com/yf0412/MRD.*

1. Introduction

Multimodal large language models (MLLMs) have driven a transformative paradigm shift in cross-modal understanding, achieving remarkable proficiency in aligning visual signals with linguistic semantics and enabling high-level cross-modal reasoning [4, 35, 44]. By bridging the gap between pixel-level visual data and abstract semantic comprehension, MLLMs have become the foundational architecture for mainstream vision tasks, from visual question answering to complex open-world scene interpretation [10, 13, 36].

However, a critical limitation remains in standard MLLM pipelines: the inadequate capability to effectively handle high-resolution images. Real-world visual data is inherently rich in fine-grained details, and the ability to perceive such features—including small objects, delicate textures, and intricate text—is essential for precise visual comprehension and complex cross-modal reasoning [38, 42, 54]. Most MLLMs exhibit inherent limitations in this regard, hampered by the constraint of fixed low input resolutions [2, 23, 24]. Recent studies indicate that existing methods remain unsatisfactory for high-resolution image tasks which is clearly demonstrated by their suboptimal performance on HR benchmarks [34, 42, 52, 56].

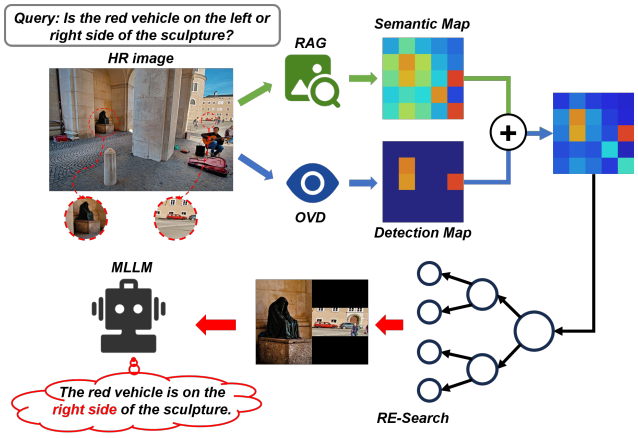


Figure 1. An illustration of the ***Multi-resolution Retrieval-Detection (MRD)*** framework for HR image understanding.

To address this limitation, enhancing HR image perception of MLLMs has emerged as a prominent research focus. The dominant "localize-and-zoom-in" paradigm improves detail perception by identifying and cropping informative sub-regions in HR images for refined processing, with existing methods falling into two categories. Training-based approaches (e.g., supervised fine-tuning (SFT) [33, 42] or reinforcement learning (RL) [55, 56]) learn effective region localization, but suffer from prohibitive computational costs, long training cycles, and poor cross-architecture transferability, limiting practical scalability. In contrast, training-free methods [18, 34, 39, 52] localize regions via attention maps or heuristic search without parameter up-

*Corresponding author.

dates. However, these top-down methods often fail in initial coarse-grained search: downsampled views lack fine-grained details, leading to irreversible error propagation, and misdirected search. Besides, they also tend to focus on main targets while neglecting others in multi-object tasks.

Inspired by the success of retrieval-augmented generation in long-context large language models (LLMs) [15], recent work Retrieval-Augmented Perception (RAP) [39] extends this method to HR image understanding. Specifically, RAP adopts an adaptive tree-search strategy and leverages a vision-based RAG model [46] to retrieve relevant image crops, guided by query-image semantic similarity and MLLM evaluation, while preserving the relative spatial layout of retrieved regions. This retrieval-based approach significantly outperforms zooming-based methods, especially in complex multi-object scenes.

Despite its promising results, retrieval-based approaches have several intrinsic limitations that severely undermine their accuracy and robustness in challenging real-world scenarios: ***1) Object Fragmentation.*** Fixed grid partitioning often splits large objects across multiple patches, destroying the holistic structural integrity of objects and producing fragmented embeddings with unreliable, biased similarity scores. ***2) Scale Sensitivity.*** Crop resolution is an intractable hyperparameter for practical deployment. Overly large patches introduce excessive background noise that dilutes fine-grained target semantic signals, while excessively small crops further exacerbate the object fragmentation problem. ***3) Background Interference.*** In HR images with textured, cluttered or crowded backgrounds, background regions frequently exhibit spurious high feature similarity to the query (e.g., from repetitive textures or shared visual statistics), generating false positives that mislead downstream MLLM inference. These failure modes collectively undermine the reliability of retrieval-augmented perception framework in complex HR image understanding tasks.

To address these challenges, we propose the ***Multi-resolution Retrieval-Detection (MRD)*** framework, a multi-scale complementary paradigm for HR image understanding, as illustrated in Fig. 1. MRD systematically boosts retrieval accuracy and robustness via joint modeling of local semantic integrity and global spatial grounding within a unified architecture. At the local scale, we introduce the *Multi-resolution Semantic Fusion* module to adapt to target objects of diverse scales. Unlike conventional fixed patch-scale strategies, we compute semantic similarity across multiple proportional resolutions and enforce cross-scale consistency via fusion. This design effectively calibrates semantic bias, alleviates object fragmentation, and preserves local semantic integrity for precise target identification. At the global scale, we develop an *Open-vocabulary Detector Enhancement* module for explicit global spatial grounding. As semantic similarity matching alone lacks sufficient spatial awareness for accurate target localization, we integrate the state-of-the-art open-vocabulary detector LLMDet [11] as a robust global prior. Leveraging the in-context learning capability of LLMs, we extract target concepts from user queries to guide the generation of a target-aware confidence map, effectively suppressing false-positive background interference. Finally, we augment multi-resolution semantic similarity with object detection confidence. This synergistic fusion amplifies responses of genuine target regions, enabling faster and more accurate relevant region localization during retrieval, and in turn guides MLLMs toward more reliable inference. Overall, MRD offers novel insights into joint modeling of local semantic integrity and global spatial grounding, advancing the frontier of HR image understanding. Extensive experiments demonstrate our superior performance and strong generalization across multiple mainstream benchmarks and diverse MLLM backbones. Our key contributions are summarized as follows:

- We propose ***MRD***, a training-free unified multi-scale framework for high-resolution image understanding, which improves retrieval accuracy via jointly modeling local semantic integrity and global spatial grounding.
- We design two core innovations for MRD: multi-resolution semantic fusion to address single-scale retrieval limitations, and open-vocabulary detector enhancement to complement semantic similarity for precise spatial localization and background suppression.
- Comprehensive experiments across diverse benchmarks and MLLMs validate the superior performance and robustness of our method, providing new insights for advancing high-resolution image understanding.

2. Related Works

2.1. Multimodal Large Language Models

Multimodal large language models have emerged as powerful foundation models, driving rapid progress in cross-modal understanding and generation across diverse vision-language tasks [44, 51]. A canonical MLLM architecture adopts a pre-trained vision encoder to map input images into visual tokens [32, 50], which are then projected into the embedding space of the backbone Large Language Model (LLM) via a multimodal connector, and jointly processed with textual inputs [8, 22, 23]. Equipped with robust cross-modal alignment and emergent reasoning capabilities, MLLMs have become the core building block for a wide spectrum of multimodal applications, spanning medical image analysis [17, 49], video understanding [40, 41], autonomous driving [14, 43], personalized assistants [1, 31], and robotic manipulation [19, 25]. Despite these impressive advances, enabling MLLMs to achieve efficient, fine-grained understanding of high-resolution images remains a critical open challenge.

2.2. High-Resolution Image Understanding

Extensive research efforts have been dedicated to advancing the HR image understanding capabilities of MLLMs, with mainstream works falling into several directions. The first line of work focuses on architectural improvement for MLLMs. These methods either design HR-specialized visual encoders [12, 21, 27, 29], or adopt dynamic resolution strategies for patch-based encoding [6, 9, 16, 53, 58], to preserve fine-grained visual details in HR inputs. Another dominant paradigm is the localize-and-zoom-in framework. Methods under this paradigm leverage SFT [33, 42] or RL [20, 37, 55, 56] to train MLLMs to actively locate relevant regions and filter redundant background information. Despite their intuitive design, these methods suffer from prohibitive training costs, long convergence cycles, and poor cross-architecture transferability, which severely limit their scalability and real-world deployment [7, 48].

Alternatively, a large body of training-free methods also follow the localize-and-zoom-in paradigm to enhance the HR understanding capability of off-the-shelf MLLMs. Works like TextCoT [28] and Zoom-Refine [47] directly leverage MLLMs' inherent reasoning to locate target regions, while ZoomEye [34] and Dyfo [18] adopt tree-search to progressively narrow down relevant regions. Though effective on single-object tasks, they frequently miss targets in complex multi-object scenarios. Another branch of training-free works [26, 52, 57] uses MLLMs' internal attention maps to guide region localization, but their performance heavily relies on the capacity of the pre-trained MLLM backbone. Recently, Retrieval-Augmented Perception [39] integrates visual retrieval-augmented generation with tree-search to retrieve semantically relevant image patches, achieving state-of-the-art performance on both single and multi-object HR understanding tasks. Nevertheless, it still suffers from inherent limitations including object fragmentation and background interference.

3. Preliminary

In this section, we first conduct an in-depth analysis of the failure cases of retrieval-based method to elucidate the inherent limitations of these approach. We then present an investigation into the correlation between the resolution of image crops and the performance of MLLMs. Building on the key insights derived from these empirical analyses, we introduce our ***MRD*** framework in Sec. 4.

3.1. Semantic Similarity

This section presents the pipeline for integrating retrieval-augmented generation into MLLMs to calculate the semantic similarity between the query embedding and image crops from HR images. Given an HR image, we first partition it collection of image crops, denoted as $P = \{p_1, p_2, \ldots, p_n\}$, where n is the total number of image crops. Following VisRAG [46], the query and each image crop are encoded respectively using the text and image encoders of a Vision-Language Model (VLM), producing a sequence of hidden representations. The similarity score $s(q, p_i)$ between the query embedding and the i-th image crop embeddings can be calculated by the cosine similarity:

$$s(q, p_i) = \frac{1}{2} \cdot (1 + \frac{q \cdot p_i}{\|q\| \cdot \|p_i\|}) \tag{1}$$

Based on these scores, the K most relevant image crops are selected and provided to the MLLM to support detailed understanding of the high-resolution input.

3.2. Limitations of Retrieval-Based Methods

To systematically characterize the inherent limitations of retrieval-based HR image understanding methods, we conduct a fine-grained case-by-case analysis of failure cases produced by the RAP [39] method on the *V* Bench* [42] with LLaVA-ov [16]. We manually categorize the root causes of each failure based on the MLLM's reasoning outputs, retrieved results, and semantic similarity scores, and define three core failure modes: ***1) Object Fragmentation (FRAG)***: Semantic bias arises from target object fragmentation, where the retrieval result only retains partial crops of the target object (Fig. 2, top). ***2) Background Interference (BG)***: Retrieved results are either dominated by or exclusively consist of high-similarity false-positive crops from irrelevant background regions (Fig. 2, bottom); (3) Other causes (e.g., inherent limitations of the MLLM itself). Notably, a single failure case can be attributed to multiple failure modes simultaneously. We quantify the impact of each failure mode by calculating the proportion of affected samples across the full dataset, with results summarized in the first row of Tab. 2. For the RAP framework , ***FRAG*** and ***BG*** errors affect 10.7% and 8.9% of the full dataset, accounting for 65.2% and 54.3% of all failure cases, respectively. Most notably, 89.6% of all failure cases are associated with either ***FRAG*** or ***BG*** errors, confirming that these two failure modes are the dominant bottlenecks restricting the performance of retrieval-based HR image understanding methods.

3.3. Impact of Image Crop Resolution

To validate the influence of crop resolution, we conduct comprehensive experiments on the *V* Bench* with the RAP method, adopting image crops of varying resolutions across a series of mainstream MLLMs; the results are summarized in Fig. 3. We observe that the selection of crop resolution imposes a pronounced impact on MLLMs' performance on HR image understanding. Specifically, while a crop resolution of 112 yields the best performance on both single and multi-object tasks, it suffers from degraded recognition accuracy compared to alternative resolutions in several cases,

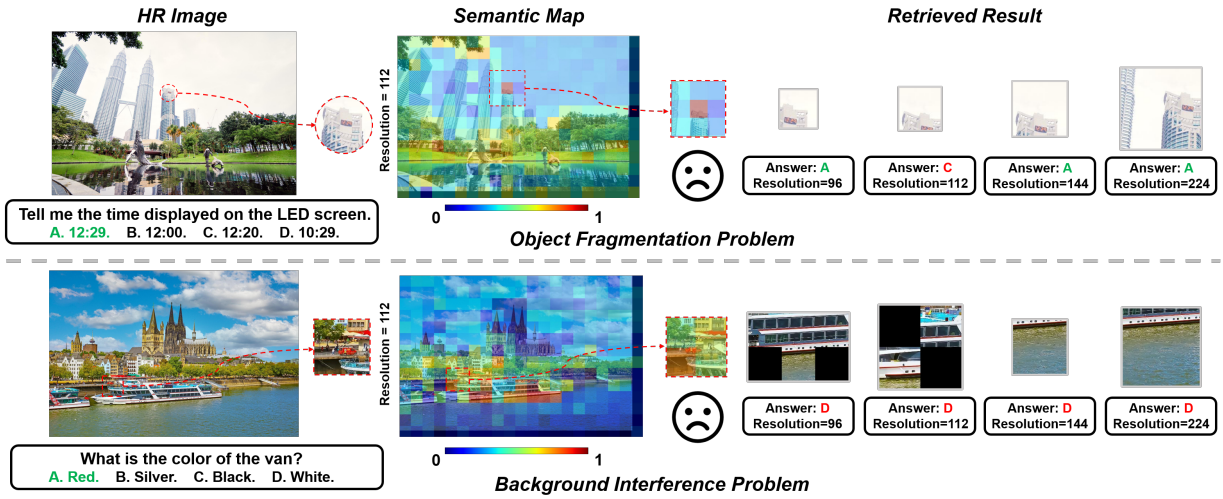


Figure 2. Object fragmentation and background interference problems in retrieval-based methods

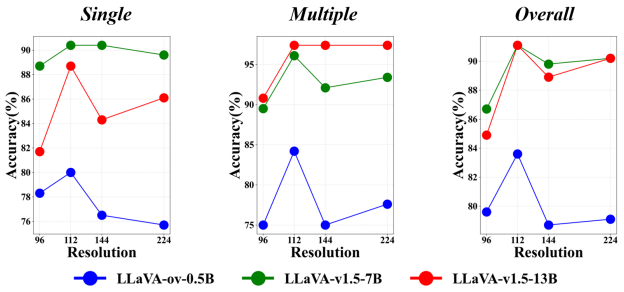


Figure 3. The impact of the resolution of retrieved image crops on MLLMs performance.

which is caused by the object fragmentation problem (Fig. 2 top). This key observation motivates us that partitioning HR images with adaptive crop resolutions, customized for target objects with diverse sizes and spatial locations, can be a more effective paradigm for HR visual understanding.

4. Method

Building on aforementioned the analysis and findings in Sec. 3, we propose a novel framework, ***Multi-resolution Retrieval Detection (MRD)***. The core of ***MRD*** lies in its multi-scale multi-resolution design, which achieves more precise target localization, suppresses background interference while preserving target semantic integrity to improve MLLMs perceptual performance for robust HR image understanding. Driven by insights in Sec. 3.3, we first design a simple yet effective *Multi-resolution Semantic Fusion* module, which adapts to target objects of varying sizes and maintains the semantic integrity of objects at the local scale. On this basis, we propose an *Open-vocabulary Detector Enhancement* module, which incorporates an open-vocabulary detection model to traverse the entire HR image via a sliding window approach at the global scale, enabling more accurate and direct localization of target regions while eliminating irrelevant background distractions. The synergistic integration of these two modules enables more accurate and efficient localization of target regions during the retrieval process, while simultaneously eliminating interference from irrelevant background. We elaborate on the complete algorithmic details and technical implementation of each component in the subsequent sections.

4.1. Multi-resolution Semantic Fusion

In Sec. 3.3, we empirically observe that image crops at different resolutions are suited for capturing objects of varying scales across scenarios. Compared to semantic similarity maps generated from a single resolution, maps derived from complementary resolutions exhibit distinct and mutually beneficial strengths. To this end, we introduce a *Multi-resolution Semantic Fusion* module to aggregate cross-resolution semantic information, as illustrated in the upper branch of Fig. 4. Specifically, we split the input high-resolution image into two sets of image crops with proportional resolutions. We denote the base resolution of the low-resolution (LR) crop set as l, and the resolution of the HR crop set as $\hat{l} = k \cdot l$, where k is a positive integer scaling factor. Let $\hat{P} = \{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m\}$ denote the set of HR crops and $P = \{p_1, p_2, \ldots, p_n\}$ denote the set of LR crops. Owing to the proportional scaling between the two resolutions, we have $n = k^2 \cdot m$ meaning each HR crop $\hat{p}_i$ spatially corresponds to exactly k^2 LR crops arranged in a $k \times k$ grid:

$$\hat{p}_i = \begin{bmatrix} \tilde{p}_{i,1} & \cdots & \tilde{p}_{i,k} \\ \vdots & \ddots & \vdots \\ \tilde{p}_{i,(k(k-1)+1)} & \cdots & \tilde{p}_{i,k^2} \end{bmatrix} \tag{2}$$

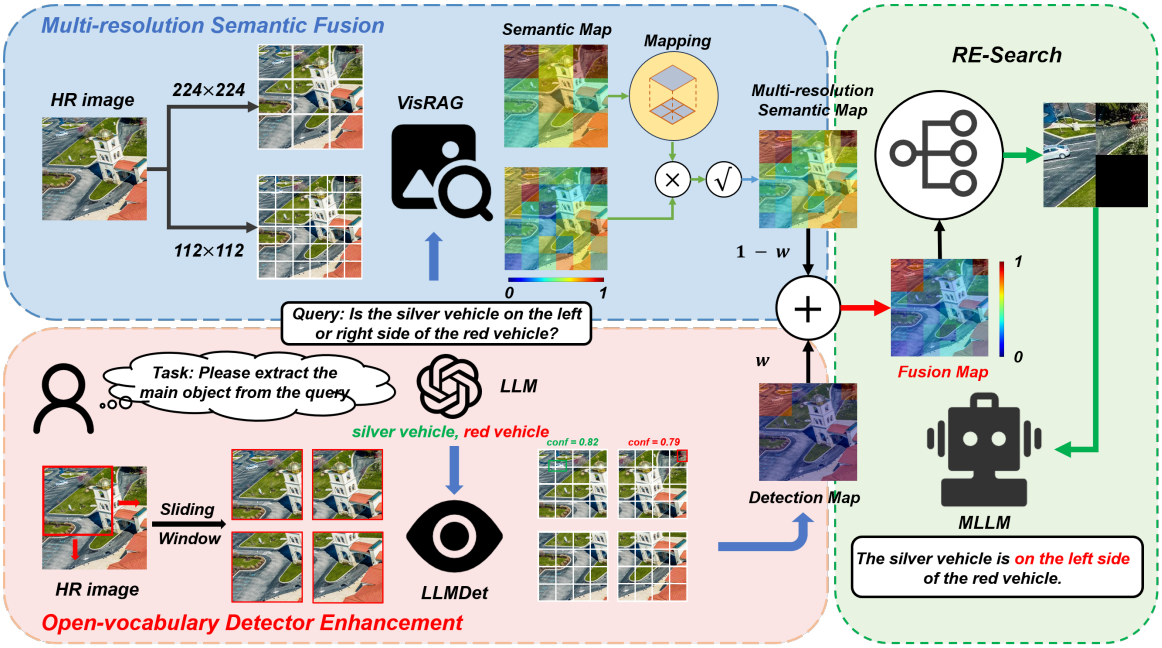


Figure 4. Overview of the proposed ***MRD*** framework. The framework generates a multi-resolution semantic similarity map via VisRAG from multi-scale crops, and a global detection confidence map via the open-vocabulary detector LLMDet with sliding-window target localization. The two maps are linearly fused, with the fused map guiding target-aware crop selection for subsequent search.

where each element $\tilde{p}_{i,t}$ corresponding to a unique LR crop $p_j \in P, i \in \{1, 2, \ldots, m\}, \quad t \in \{1, 2, \ldots, k^2\}$.

Using the cosine similarity function defined in Eq. (1), we compute the semantic similarity between the query embedding and crops embedding, yielding two sets of similarity scores: $\hat{S} = \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_m\}$ for the HR crop set, and $S = \{s_1, s_2, \ldots, s_n\}$ for the LR crop set, formalized as:

$$\hat{s}_i = s(f(q), g(\hat{p}_i)), \quad s_j = s(f(q), g(p_j)) \tag{3}$$

where $f(\cdot)$ and $g(\cdot)$ denote the text encoder for the input query and the image encoder for image crops, respectively, with $i \in \{1, 2, \ldots, m\}$ and $j \in \{1, 2, \ldots, n\}$.

Leveraging the spatial correspondence between HR and LR crops defined in Eq. (2), we project each HR similarity score $\hat{s}_i$ to its corresponding k^2 LR crop positions, replicating the score across the associated $k \times k$ LR grid. This projection operation is written as:

$$\tilde{S} = H(\hat{S}) \tag{4}$$

where $\tilde{S} = \{\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_n\}$ is the projected HR similarity set, which shares the exact spatial dimensions and element count as the LR similarity set S, ensuring perfect crop-wise alignment between the two. With the spatially aligned similarity sets $\tilde{S}$ and S, we perform element-wise consistency fusion to generate the final multi-resolution semantic similarity scores $S^m = \{s_1^m, s_2^m, \ldots, s_n^m\}$, computed as:

$$s_t^{\mathrm{m}} = \sqrt{\tilde{s}_t \cdot s_t}, \quad t \in 1, 2, \ldots, n \tag{5}$$

Finally, we reshape the 1D fused similarity set S^m into a 2D semantic similarity map $s^{\mathrm{m}}(i, j)$ of size $H \times W$, where H and W denote the height and width of the LR crop grid respectively. Fusing cross-resolution semantic similarities enables correction of LR similarity degradation caused by complete objects being split across multiple LR crops. Our method reinforces consistent similarity responses across all parts of the target object, preserving its semantic integrity in subsequent retrieval steps.

4.2. Open-vocabulary Detector Enhancement

To alleviate background interference in the multi-resolution semantic similarity map, we propose an *Open-vocabulary Detector Enhancement* module. This module introduces explicit object-level localization constraints using the open-vocabulary detector LLMDet, paired with a crop-aligned sliding-window strategy. First, we harness the in-context learning capability of LLMs to extract core target entities from the input query, which serve as the target categories for the open-vocabulary detector. To achieve precise alignment with the semantic similarity map generated by our multi-resolution fusion module, we map the detection confidence scores of target bounding boxes to their corresponding image crops, yielding a detection confidence map that quantifies the target presence probability for each crop. This map delivers explicit, spatially accurate localization guidance that is inherently complementary to our fine-grained semantic similarity map. The details are discussed below.

Object Extraction. To support open-vocabulary detection without pre-defined category limits, we first use an LLM to dynamically extract target object entities from free-form textual queries. Given an input query Q, we adopt in-context learning to extract primary target entities, which form the detection category set for our framework. Formally, the extraction process is defined as:

$$O = \text{LLM}(\mathcal{P}\text{system}, \mathcal{E}\text{examples}, Q) \tag{6}$$

where O denotes the set of extracted target objects, $\mathcal{P}$system is the system prompt specifying extraction rules, and $\mathcal{E}$examples contains few-shot demonstration examples for in-context learning.

Sliding-window Object Detection. To handle ultra-high-resolution images that exceed the detector's optimal input size, we adopt a sliding-window detection strategy which is strictly spatially aligned with the $H \times W$ non-overlapping crop grid used in our multi-resolution semantic fusion module (with total crop count $n = H \times W$).

Specifically, a sliding window of size $h \times w$ crops (where $h < H$ and $w < W$) traverses the entire HR image with a pre-defined stride, yielding T sliding windows denoted as $\Omega = \{w_1, w_2, ..., w_T\}$. For each sliding window w_t we use LLMDet to perform open-vocabulary detection with the extracted target objects O. The detector outputs a set of bounding boxes $\mathcal{B}_t = \{b_1, b_2, \ldots, b_{K_t}\}$ and corresponding confidence scores c_k, where c_k indicates the confidence that bounding box b_k contains the target object.

We first apply a confidence threshold τ to filter out low-quality detections:

$$\mathcal{B}_t^{\text{filter}} = \{b_k \in \mathcal{B}_t \mid c_k > \tau\} \tag{7}$$

Subsequently, we generate a window detection confidence map $\mathbf{c}_t^w \in \mathbb{R}^{h \times w}$ for each sliding window. For a crop at local coordinate (p, q) within the window, we assign the maximum confidence score of all filtered bounding boxes that cover this crop; the score is set to 0 if no valid box covers the crop:

$$c_t^{\text{w}}(p, q) = \max_{b_k \in \mathcal{B}_t^{\text{filter}}} \{c_k \cdot \mathbb{I}[(p, q) \in b_k]\} \tag{8}$$

where $\mathbb{I}[\cdot]$ is the indicator function, which equals 1 if the crop at (p, q) lies within bounding box b_k, and 0 otherwise.

To aggregate local window into a global unified detection confidence map $\mathbf{c}^{\text{g}} \in \mathbb{R}^{H \times W}$ (aligned with the semantic similarity map), we adopt an average fusion strategy to mitigate boundary artifacts from overlapping windows and smooth single-window detection noise. For a crop at global coordinate (i, j) in the HR image grid, we denote $\mathcal{T}_{i,j}$ as the set of sliding windows that contain this crop, and (t_i, t_j) as its corresponding local coordinate in the t-th sliding window w_t. The final global confidence score is computed as the average of all valid local scores assigned to this crop:

$$c^{\text{g}}(i, j) = \frac{1}{|\mathcal{T}_{i,j}|} \sum_{t \in \mathcal{T}_{i,j}} c_t^{\text{w}}(t_i, t_j) \tag{9}$$

where $i \in \{1, 2, \ldots, H\}$ and $j \in \{1, 2, \ldots, W\}$.

Semantic-Detection Fusion. The global detection confidence map provides explicit target localization and strong background suppression, but lacks the ability to capture fine-grained semantic differences within the target object. In contrast, our multi-resolution semantic similarity map encodes fine-grained query-image matching, but is vulnerable to background clutter and boundary ambiguity. To leverage the complementary strengths of the two branches, we fuse them via a linear combination to generate the final fused similarity map:

$$s^{\text{f}}(i, j) = (1 - w) \cdot s^{\text{m}}(i, j) + w \cdot c^{\text{g}}(i, j) \tag{10}$$

where w is a balancing weight. This synergistic fusion strategy simultaneously achieves precise target localization and full semantic integrity preservation, which in turn enables more efficient and accurate extraction and retrieval of key regions during the subsequent search process. The detailed description of the follow-up *Retrieved-Exploration Search* pipeline is provided in [39].

5. Experiments

5.1. Experimental Settings

Benchmarks. We evaluate our *MRD* on two HR image benchmarks: *V* Bench* [42] and *HR-Bench* [38] with *4K* or *8K* resolution. Both cover single-object and multi-object understanding tasks for HR images.

Baseline Methods. We evaluate our *MRD* against three groups of baselines on HR benchmarks: 1) mainstream open-source MLLMs, including LLaVA-v1.6 series [24], LLaVA-HR-X [30], InternVL-1.5 [5] and Yi-VL [45]; 2) closed-source MLLMs GPT-4o [13] and Qwen-VL-max [3] as upper-bound references; 3) SOTA high-resolution specialized methods Zoom Eye [34] and RAP [39], which share the same LLaVA-v1.5-7B [23] and LLaVA-ov-0.5B [16] backbones with our approach for fair comparison.

5.2. Main Results

As shown in Tab. 1, our proposed *MRD* framework consistently achieves significant performance gains across all high-resolution benchmarks, sub-tasks, and model configurations, outperforming both vanilla MLLM baselines and state-of-the-art high-resolution perception methods. Built on LLaVA-v1.5-7B, MRD delivers a 46.9% absolute improvement in overall accuracy on *V* Bench* (nearly doubling the baseline), along with up to 23.8% and 22.7% absolute overall gains on *HR-Bench 4K* and *HR-Bench 8K*, re-

Table 1. Comparison of our ***MRD*** with existing works on high-resolution benchmarks

Method	*V* Bench*			*HR-Bench 4K*			*HR-Bench 8K*		
	Attribute	***Spatial***	***Overall***	***FSP***	***FCP***	***Overall***	***FSP***	***FCP***	***Overall***
Open-source MLLMs									
LLaVA-v1.6-7B [24]	60.9	63.2	61.8	49.0	46.8	47.9	37.3	44.3	40.8
LLaVA-v1.6-13B [24]	60.0	64.5	61.8	49.8	41.3	45.5	38.0	38.3	38.1
LLaVA-v1.6-34B [24]	-	-	-	55.3	50.5	52.9	44.5	50.3	47.4
LLaVA-HR-X-13B [30]	-	-	-	61.3	46.0	53.6	49.5	44.3	46.9
LLaVA-HR-X-7B [30]	51.3	64.5	56.5	57.8	46.3	52.0	42.0	41.3	41.6
InternVl-1.5-26B [5]	-	-	-	69.5	51.8	60.6	69.3	48.5	57.9
Yi-VL-34B [45]	-	-	-	46.0	42.8	44.4	39.5	38.5	39.0
Closed-source MLLMs									
GPT-4o [13]	-	-	66.0	70.0	48.0	59.0	62.0	49.0	55.5
Qwen-VL-max [3]	-	-	-	65.0	**52.0**	58.5	54.0	**51.0**	52.5
Baselines and MRD									
LLaVA-v1.5-7B [23]	43.5	56.6	48.7	38.5	33.8	36.1	33.0	31.3	32.1
LLaVA-v1.5-7B-Zoom Eye [34]	83.5	82.9	83.3	67.8	38.8	53.3	65.5	36.0	50.8
LLaVA-v1.5-7B-RAP [39]	90.4	**96.1**	91.1	73.8	40.5	57.1	72.3	35.3	53.8
LLaVA-v1.5-7B-MRD *(ours)*	**97.4**	**96.1**	**95.6**	76.8	42.7	59.9	73.0	36.5	54.8
LLaVA-ov-0.5B [16]	63.5	64.5	63.9	63.5	39.5	51.5	47.3	38.3	42.8
LLaVA-ov-0.5B-Zoom Eye [34]	85.2	73.7	80.6	75.5	39.8	57.6	68.5	38.3	53.4
LLaVA-ov-0.5B-RAP [39]	80.0	84.2	83.6	80.3	42.3	61.3	**81.8**	45.3	63.5
LLaVA-ov-0.5B-MRD *(ours)*	89.6	85.6	88.9	**84.0**	45.2	**64.6**	**81.8**	47.3	**64.5**

spectively. It consistently surpasses the SOTA RAP method across all datasets and model settings, with an average overall improvement of 2.8%. MRD brings stable performance boosts to both single and multi-object tasks, with particularly prominent gains on single-object tasks. For the single-object attribute recognition sub-task of *V* Bench*, MRD achieves a 53.9% absolute accuracy gain over the vanilla LLaVA-v1.5-7B baseline, and a 7.0% gain over RAP. We attribute this advantage to the integrated detection module, which provides precise object localization for fine-grained visual understanding. These results validate that MRD significantly enhances the high-resolution image perception and reasoning capabilities of MLLMs.

5.3. Visualization of Module Effects

To intuitively verify the effectiveness of our core modules, we visualize the confidence map and retrieved result using either *Multi-resolution Semantic Fusion* module or the *Open-vocabulary Detector Enhancement* module in Fig. 5.

The *Multi-resolution Semantic Fusion* module obtains more accurate target information by integrating semantic similarity maps from different resolutions. As shown in the upper part of Fig. 5, this module leverages the high-resolution semantic similarity map to correct the semantic deviation caused by the splitting of the target object across multiple crops under single-resolution processing. This correction effectively alleviates the above semantic deviation and better preserves the integrity of the target object.

For the *Open-vocabulary Detector Enhancement* module, we introduce an open-vocabulary object detection model to achieve more accurate and direct localization of the target object at the global scale. As shown in the lower part of Fig. 5, the sliding-window detection results can effectively identify the location of the target object. By combining the detection results with the semantic similarity scores, our method amplifies the scores of the crops containing the target object, while suppressing the interference of false-positive crops with high semantic similarity and irrelevant background information.

5.4. Ablation Study

To better understand the contributions of different modules in our ***MRD*** framework, we conduct comprehensive quantitative ablation studies on the *V* Bench* using the LLaVA-ov-0.5B model. As shown in Tab. 2, the standalone OVD (first row) model achieves higher accuracy for single-object tasks via precise target localization and reduced background interference. However, its performance on multi-object tasks is inferior to RAP which may due to missing object in complex multiple object scene. When equipping RAP with multi-resolution semantic fusion (second row), we observe consistent performance gains on both single and

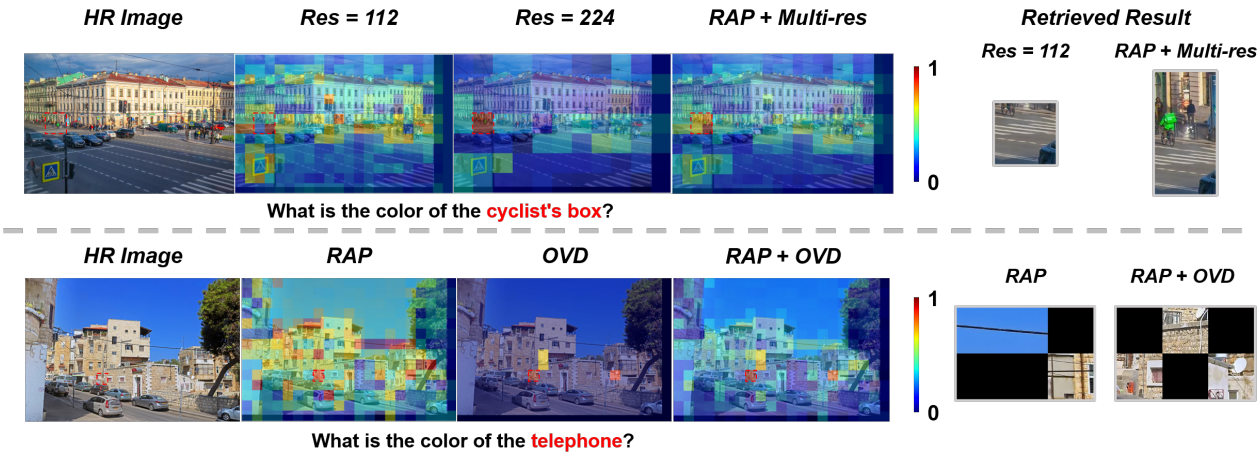


Figure 5. Visualization of the effects of proposed modules in ***MRD***. Upper: Visualization of the effects of the *Multi-resolution Semantic Fusion* module. Lower: Visualization of the effects of the *Open-vocabulary Detector Enhancement* module.

Table 2. Ablation study of different module in ***MRD***.

Method	*Accuracy*			*Error Rate*		
	Attribute (↑)	*Spatial* (↑)	*Overall* (↑)	*BG* (↓)	*FRAG* (↓)	*BG/FRAG* (↓)
RAP	80.0	84.2	83.6	10.7 (64.2%)	8.9 (54.3%)	14.7 (89.6%)
OVD	84.3 (+4.3)	82.9 (-1.3)	85.3 (+1.7)	5.7 (-46.7%)	6.2 (-30.3%)	9.3 (-36.7%)
RAP+Multi-Res	83.5 (+3.5)	**85.5** (+1.3)	85.8 (+2.2)	6.7 (-37.4%)	5.3 (-40.4%)	9.7 (-34.0%)
RAP+OVD	86.1 (+6.1)	84.2 (+0.0)	86.7 (+3.1)	4.9 (-54.2%)	5.8 (-34.8%)	8.9 (-39.5%)
All (MRD)	**89.6 (+9.6)**	**85.5 (+1.3)**	**88.9 (+5.3)**	**4.0 (-62.6%)**	**4.4 (-50.6·%)**	**7.6 (-48.3%)**

Table 3. Comparison of inference efficiency on *HR-Bench4K*.

Method	*HR-Bench4K*				
	RAG (sec)	*DET (sec)*	*Search (sec)*	*Total (sec)*	*Max Mem (GB)*
RAP (ov-0.5B)	9.6	0	55.6	67.0	18.9
MRD (ov-0.5B)	14.7 (+53.1%)	16.1	27.5 (-50.5%)	59.6 (-11.0%)	21.2 (+12.2%)
RAP (v1.5-7B)	9.8	0	52.8	63.4	21.2
MRD (v1.5-7B)	14.5 (+48.0%)	15.8	15.2 (-71.2%)	53.4 (-26.2%)	23.4 (+10.4%)

multi-object tasks, as it corrects semantic deviations, effectively mitigates object fragmentation, and preserves object completeness, making it capable of handling objects of varying sizes across diverse scenarios.

Fusing the semantic similarity map obtained from RAP with the detection confidence map from OVD (fourth row) further suppresses background interference and boosts single-object task performance. By further integrating multi-resolution semantic fusion, we achieve consistent and prominent improvements across both single- and multi-object tasks, verifying the effectiveness of our fusion strategy. The two core components are complementary and jointly necessary to address both background and fragmentation errors, ensuring consistent improvements across all settings. As a whole, our full MRD framework integrating all modules outperforms RAP by 5.3% in overall accuracy, demonstrating the superiority of our MRD framework.

5.5. Efficiency

We benchmark the average per-sample runtime and computational overhead of our ***MRD*** against the RAP method on the *HR-Bench 4K* dataset. Benefiting from our fused map that enables more accurate localization of target regions, we can drastically reduce the maximum number of search steps to 50 during the search phase while preserving full model performance. As shown in Tab. 3, although MRD introduces additional GPU memory consumption and computational cost due to multi-resolution fusion and sliding-window detection, it yields a remarkable acceleration of the search phase, which leading to a shorter overall runtime.

6. Conclusion

In this work, we present ***Multi-resolution Retrieval-Detection (MRD)***, a novel training-free method to boost MLLMs' high-resolution image understanding. MRD leverages multi-resolution semantic similarity to correct single-resolution similarity maps and preserve target object integrity, and introduces a sliding-window open-vocabulary detection (OVD) model for precise target localization. Extensive experiments on multiple HR benchmarks with various MLLMs validate MRD's effectiveness and its superior HR image understanding performance. This work provides new insights into jointly modeling local semantic integrity and global spatial grounding for HR image understanding.

Acknowledgment

This work is funded in part by the National Natural Science Foundation of China (Grant No. 62372480), GuangDong Basic and Applied Basic Research Foundation (2025A1515011361), Shenzhen Science and Technology Program (JCYJ20240813110459017).

References

[1] Ruichuan An, Sihan Yang, Renrui Zhang, Zijun Shen, Ming Lu, Gaole Dai, Hao Liang, Ziyu Guo, Shilin Yan, Yulin Luo, et al. Unictokens: Boosting personalized understanding and generation via unified concept tokens. *arXiv preprint arXiv:2505.14671*, 2025. 2

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 6, 7

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1

[5] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Sci. China Inf. Sci.*, 67(12), 2024. 6, 7

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 3

[7] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025. 3

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 2

[9] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *Advances in Neural Information Processing Systems*, 37:42566–42592, 2024. 3

[10] Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*, 2024. 1

[11] Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14987–14997, 2025. 2

[12] Chunjiang Ge, Sijie Cheng, Ziming Wang, Jiale Yuan, Yuan Gao, Jun Song, Shiji Song, Gao Huang, and Bo Zheng. Convllava: Hierarchical backbones as visual encoder for large multimodal models. *arXiv preprint arXiv:2405.15738*, 2024. 3

[13] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 6, 7

[14] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, et al. Emma: End-to-end multimodal model for autonomous driving. *Transactions on Machine Learning Research*. 2

[15] Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. Long-context llms meet rag: Overcoming challenges for long inputs in rag. *arXiv preprint arXiv:2410.05983*, 2024. 2

[16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3, 6, 7

[17] Cheng-Yi Li, Kao-Jung Chang, Cheng-Fu Yang, Hsin-Yu Wu, Wenting Chen, Hritik Bansal, Ling Chen, Yi-Ping Yang, Yu-Chun Chen, Shih-Pin Chen, et al. Towards a holistic framework for multimodal llm in 3d brain ct radiology report generation. *Nature Communications*, 16(1):2258, 2025. 2

[18] Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9098–9108, 2025. 1, 3

[19] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18061–18070, 2024. 2

[20] Xuchen Li, Xuzhao Li, Jiahui Gao, Renjie Pi, Shiyu Hu, and Wentao Zhang. Look less, reason more: Rollout-guided adaptive pixel-space reasoning. *arXiv preprint arXiv:2510.01681*, 2025. 3

[21] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 1, 2, 6, 7

[24] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024. 1, 6, 7

[25] Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Pengju An, Xiaoqi Li, Kaichen Zhou, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. *Advances in Neural Information Processing Systems*, 37:40085–40110, 2024. 2

[26] Xianjie Liu, Yiman Hu, Yixiong Zou, Liang Wu, Jian Xu, and Bo Zheng. Hide: Rethinking the zoom-in method in high resolution mllms via hierarchical decoupling. *arXiv preprint arXiv:2510.00054*, 2025. 3

[27] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 3

[28] Bozhi Luan, Hao Feng, Hong Chen, Yonghui Wang, Wengang Zhou, and Houqiang Li. Textcot: Zoom-in for enhanced multimodal text-rich image understanding. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 3

[29] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024. 3

[30] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 6, 7

[31] Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'llava: Your personalized language and vision assistant. *Advances in Neural Information Processing Systems*, 37:40913–40951, 2024. 2

[32] 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*, pages 8748–8763. PmLR, 2021. 2

[33] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024. 1, 3

[34] Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6613–6629, 2025. 1, 3, 6, 7

[35] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1

[36] LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Ankit Anand, Avishkar Bhoopchand, Brett Wiltshire, Daniel Gillick, Daniel Kasenberg, et al. Evaluating gemini in an arena for learning. *arXiv preprint arXiv:2505.24477*, 2025. 1

[37] Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025. 3

[38] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7907–7915, 2025. 1, 6

[39] Wenbin Wang, Yongcheng Jing, Liang Ding, Yingjie Wang, Li Shen, Yong Luo, Bo Du, and Dacheng Tao. Retrieval-augmented perception: High-resolution image perception meets visual rag. *arXiv preprint arXiv:2503.01222*, 2025. 1, 2, 3, 6, 7

[40] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024. 2

[41] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3272–3283, 2025. 2

[42] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024. 1, 3, 6

[43] Shuo Xing, Chengyuan Qian, Yuping Wang, Hongyuan Hua, Kexin Tian, Yang Zhou, and Zhengzhong Tu. Openemma: Open-source multimodal model for end-to-end autonomous driving. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1001–1009, 2025. 2

[44] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12): nwae403, 2024. 1, 2

[45] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng

Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai. *CoRR*, abs/2403.04652, 2024. 6, 7

[46] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024. 2, 3

[47] Xuan Yu, Dayan Guan, and Yanfeng Gu. Zoom-refine: Boosting high-resolution multimodal understanding via localized zoom and self-refinement. *arXiv preprint arXiv:2506.01663*, 2025. 3

[48] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025. 3

[49] Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16 (1):3108, 2025. 2

[50] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2

[51] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12401–12430, 2024. 2

[52] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. *arXiv preprint arXiv:2502.17422*, 2025. 1, 3

[53] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024. 3

[54] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024. 1

[55] Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025. 1, 3

[56] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. 1, 3

[57] Liangyu Zhong, Fabio Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. Focus: Internal mllm representations for efficient fine-grained visual question answering. *arXiv preprint arXiv:2506.21710*, 2025. 3

[58] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 3

MRD: Multi-resolution Retrieval-Detection Fusion for High-Resolution Image Understanding

Supplementary Material

A. Implement Details of *MRD*

Given an input high-resolution image *I*, we first partition it into fixed-size local crops, with crop dimensions predefined according to the native resolution of the input image, following the protocol of *RAP* [39]. Specifically for our proposed ***MRD*** framework, we set the crop resolution to 112, 224, and 448 for the V^* *Bench*, *HR-Bench-4K*, and *HR-Bench-8K* benchmarks, respectively, to match the native resolution of each dataset. For the multi-resolution semantic fusion module in our framework, we fix the resolution ratio between the high-resolution and low-resolution branches to $k = 2$ for all experiments.

For the sliding window detection pipeline, we configure the window size and stride to strike a balance between inference efficiency and detection accuracy: we set the (window size, stride) pairs to (1232, 896) for V^* *Bench*, (2240, 1792) for *HR-Bench-4K*, and (3136, 2688) for *HR-Bench-8K*, respectively. We set the default detection confidence threshold to 0.3 to filter out low-quality bounding box predictions, and fix the weight w of the detection confidence map to 0.4 for the semantic detection map fusion step. For the subsequent *Retrieval-Exploration Search* pipeline, we adopt the same hyperparameter settings as the *RAP* baseline method, with one critical exception: our method achieves peak performance with fewer search iterations, as validated in Sec. B.3. Accordingly, we fix the maximum number of search steps to 50 across all our experiments.

For all subsequent hyperparameter ablation studies, all other components of our proposed ***MRD*** framework are fixed to the above default settings unless explicitly specified otherwise.

B. Additional Ablation Studies

To further understand the behavior of our framework and evaluate the robustness of the proposed ***MRD*** module, we conduct additional ablation studies by varying several key hyperparameters. These experiments aim to analyze how different parameter settings influence the overall performance of the system.

Specifically, we vary important hyperparameters such as *crop resolution*, *more resolution fusion*, *maximum search steps*, *detection weight*, sliding window size, and detection confidence threshold (Sec. B.1 to Sec. B.1). For each experiment, we modify only one hyperparameter while keeping all others fixed, allowing us to isolate the effect of each factor on model performance.

All experiments are conducted on the V^* *Bench* using two representative multimodal large language models, LLaVA-ov-0.5B and LLaVA-v1.5-7B. This setup enables us to evaluate the sensitivity of the proposed method across different model scales. Unless otherwise specified, all remaining hyperparameters follow the default settings described in Sec. A. The detailed results are presented in the following subsections.

B.1. Effect of Crop Resolution

To systematically investigate the impact of crop resolution on performance, we conduct ablation experiments under a range of crop size settings. The results are shown in Fig. 6. For the single-object task (Fig. 6 (a)), the proposed ***MRD*** demonstrates consistently stable performance across all tested resolutions for both backbones, with only minor fluctuations. In contrast, the *RAP* baseline exhibits noticeable performance instability under varying crop resolutions, particularly when using the lightweight LLaVA-ov-0.5B model. For the multi-object task (Fig. 6 (b)), the performance gap between ***MRD*** and *RAP* becomes smaller when using the stronger LLaVA-v1.5-7B backbone. However, when switching to the smaller LLaVA-ov-0.5B model, ***MRD*** still maintains stable performance across different resolutions, further demonstrating its robustness to resolution changes. Overall, ***MRD*** consistently outperforms the *RAP* baseline across all tested crop resolutions and backbone configurations.

These results suggest our ***MRD*** effectively mitigate the object fragmentation issue that often arises when objects are split across adjacent crops under different resolutions. As a result, our framework shows reduced sensitivity to crop resolution and achieves more stable performance, particularly in the single-object setting.

B.2. Effect of More Resolution Fusion

To further investigate the impact of using additional resolution scales in the *Multi-resolution Semantic Fusion* module, we conduct experiments by incorporating multi-scale fusion into the RAP framework. Specifically, we evaluate three scale settings with $k = 1, 2, 4$, while keeping the crop resolution fixed at 112 for all image patches. Experiments are conducted on the V^* *Bench* using two representative backbones, LLaVA-ov-0.5B and LLaVA-v1.5-7B. The results are summarized in Tab. 4 For the LLaVA-ov-0.5B model, fusing three resolution scales ($k = 1, 2, 4$) achieves better performance than using two scales or a single scale

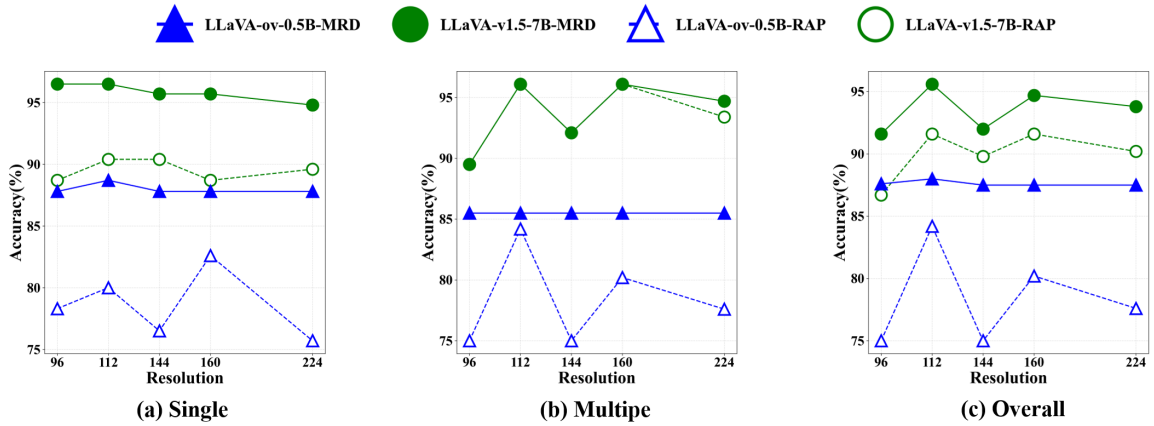


Figure 6. The effect of the resolution of image crops on MLLMs performance. Single and Multiple represent the attribute recognition and spatial reasoning tasks in V^* Bench. (a) Single-object Task. (b) Multi-object Task. (c) Overall Performance.

.

for both the single-object and multi-object tasks. In contrast, for the LLaVA-v1.5-7B backbone, using two scales ($k = 1, 2$) leads to better performance than incorporating an additional resolution level.

These results suggest that introducing more resolution scales does not necessarily lead to consistent performance improvements across different backbone models. However, multi-resolution fusion consistently outperforms the single-resolution setting in all cases, indicating the effectiveness of leveraging complementary information from different crop scales. Considering both performance and computational efficiency, we adopt a two-resolution scheme ($k = 1, 2$) in practice.

Table 4. Effect of more resolution fusion on $V*$ *Bench.*

Method	*V* Bench*		
	Attribute	Spatial	Overall
RAP (LLaVA-ov-0.5B)	80.0	84.2	83.6
+Multi-Res (k=1,2)	**83.5**	85.5	85.8
+Multi-Res (k=1,4)	80.0	86.8	84.4
+Multi-Res (k=1,2,4)	**83.5**	**88.2**	**86.7**
RAP (LLaVA-v1.5-7B)	90.4	**96.1**	91.1
+Multi-Res (k=1,2)	**94.8**	**96.1**	**94.2**
+Multi-Res (k=1,4)	93.9	94.7	93.3
+Multi-Res (k=1,2,4)	93.9	**96.1**	**94.2**

B.3. Effect of Maximum Search Steps

The performance of ***MRD*** and *RAP* under different maximum search steps is illustrated in Fig. 7.For the single-object task (Fig. 7 (a)), ***MRD*** consistently outperforms *RAP* across all tested maximum step settings on both the LLaVA-ov-0.5B and LLaVA-v1.5-7B backbones, demonstrating stable performance under varying search budgets. For the multi-object task (Fig. 7 (b)), ***MRD*** is slightly inferior to *RAP* only when using the LLaVA-v1.5-7B backbone with a very small number of maximum steps. As the maximum step increases, ***MRD*** quickly surpasses *RAP* and maintains superior performance. Overall, ***MRD*** achieves better results than *RAP* across most settings. Notably, ***MRD*** with the lightweight LLaVA-ov-0.5B backbone achieves performance that is only marginally lower than *RAP* with the much larger LLaVA-v1.5-7B model.

Another important observation is that ***MRD*** reaches its peak performance with a relatively small number of maximum search steps (Max Step = 50). This indicates that our method can achieve strong performance with a limited search budget, reducing search time while maintaining high accuracy in practical scenarios.

B.4. Effect of Detection Weight

The results under different detection weight settings are presented in Fig. 8. We observe that relying solely on the multi-resolution semantic similarity map (weight w= 0) or solely on the detection map (weight w = 1) does not yield optimal performance for either task. In contrast, combining the two maps leads to improved results, indicating that the semantic similarity map and the detection map provide complementary information.

Overall, the optimal detection weight varies slightly across different backbones. The lightweight LLaVA-ov-0.5B model achieves its best performance at a detection weight of 0.4, while LLaVA-v1.5-7B performs best when the detection weight is set to 0.2. These observations further validate the benefit of integrating both semantic and

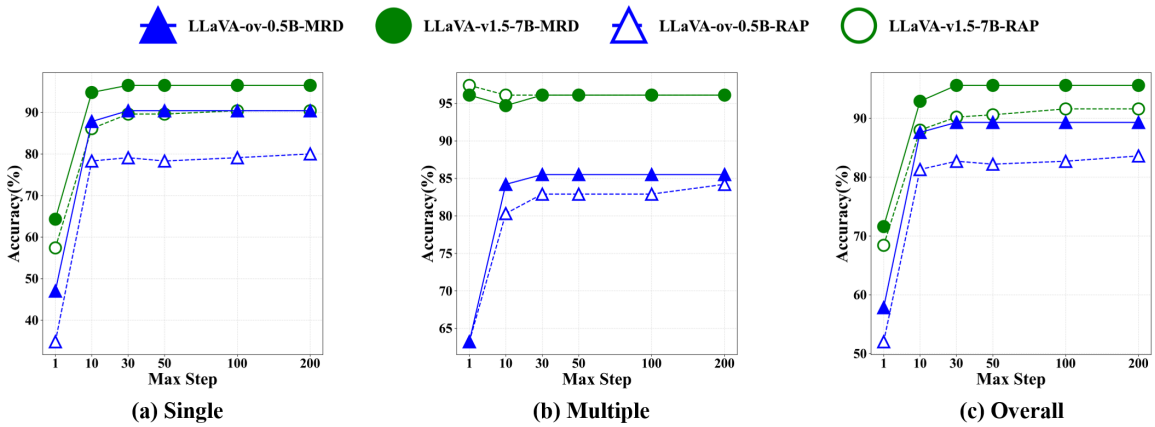


Figure 7. The effect of the maximum search steps in ***MRD*** and *RAP*.

detection cues in our framework.

B.5. Effect of Window Size

As shown in Fig. 9, the choice of sliding-window size for object detection also has a noticeable impact on the final performance. In most settings, using a smaller sliding-window size (Window Size = 896) leads to slightly better results, except for the multi-object task with the LLaVA-ov-0.5B backbone. This improvement can be attributed to the reduced background context within smaller windows, which helps suppress irrelevant regions and enables the detector to focus more precisely on the target objects.

However, reducing the window size also introduces additional computational costs. Specifically, smaller windows require a larger number of sliding windows to cover the entire high-resolution image, which increases both the number of detection operations and the overall processing time. As a result, although smaller windows may provide marginal accuracy gains, they lead to reduced efficiency in large-scale inference scenarios.

To balance detection accuracy and computational efficiency, we adopt a larger sliding-window size (Window Size = 1232) as the default configuration in our framework. This setting provides a favorable trade-off between performance and runtime, while still maintaining competitive detection accuracy across different tasks and backbone models.

B.6. Effect of Detection Confidence Threshold

The effect of different detection confidence thresholds is illustrated in Fig. 10. We evaluate a range of threshold values to analyze how filtering detection candidates influences the overall performance. From Fig. 10, we observe that the best performance is consistently achieved when the confidence threshold is set to 0.3. This trend holds across both backbone models (LLaVA-ov-0.5B and LLaVA-v1.5-7B) and across different tasks, including the single-object task, the multi-object task, and the overall evaluation.

As the threshold increases, the performance gradually decreases in most cases. A higher threshold filters out more detection candidates, which may remove useful object proposals and reduce the coverage of potential target regions. In contrast, using a lower threshold retains more candidate regions, allowing the subsequent reasoning and fusion modules to better identify relevant objects.

Based on these observations, we adopt a confidence threshold of 0.3 as the default setting in our framework, as it consistently provides the best performance across different tasks and backbone models.

C. More Visualization Results

C.1. Examples of Single-object Perception Task

Fig. 11 presents two qualitative examples of the single-object perception task from each HR benchmark, comparing ***MRD*** with *RAP* using the LLaVA-v1.5-7B backbone. From left to right, we visualize the HR image, the semantic similarity map produced by *RAP*, the object detection confidence map, the semantic–detection fusion map generated by ***MRD***, and the final predictions from *RAP* and ***MRD***. From the visualization of the *RAP* semantic similarity maps, it can be observed that crop-based partitioning may split a complete object across multiple image crops, resulting in inconsistent semantic similarity scores across different parts of the object. Such inconsistencies can negatively affect the subsequent retrieval process. For example, in the second case from *HR-Bench4K*, *RAP* retrieves only the right half of the speed-limit sign, leading to an incorrect final prediction. In addition, the similarity maps often contain false positives. For instance, in the first example from *HR-Bench8K*, the sky region—although irrelevant to

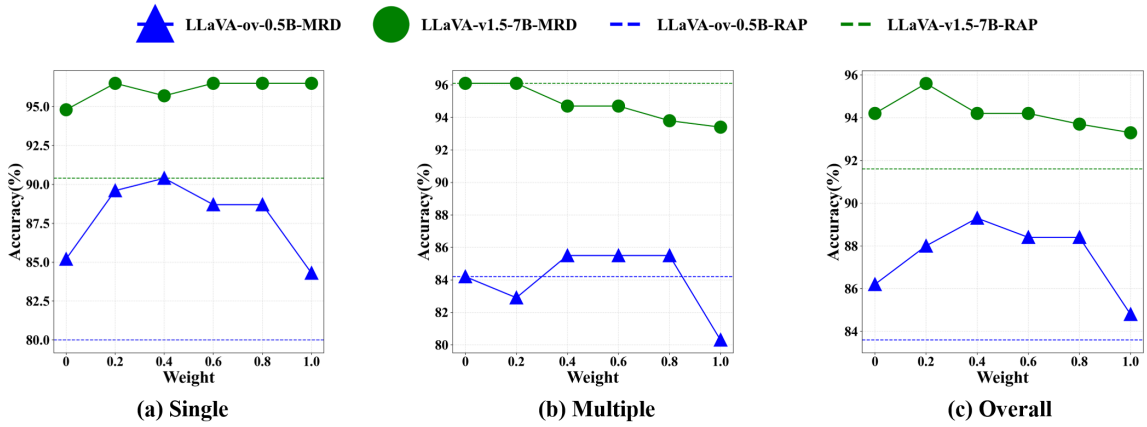


Figure 8. The effect of the detection weight in ***MRD***.

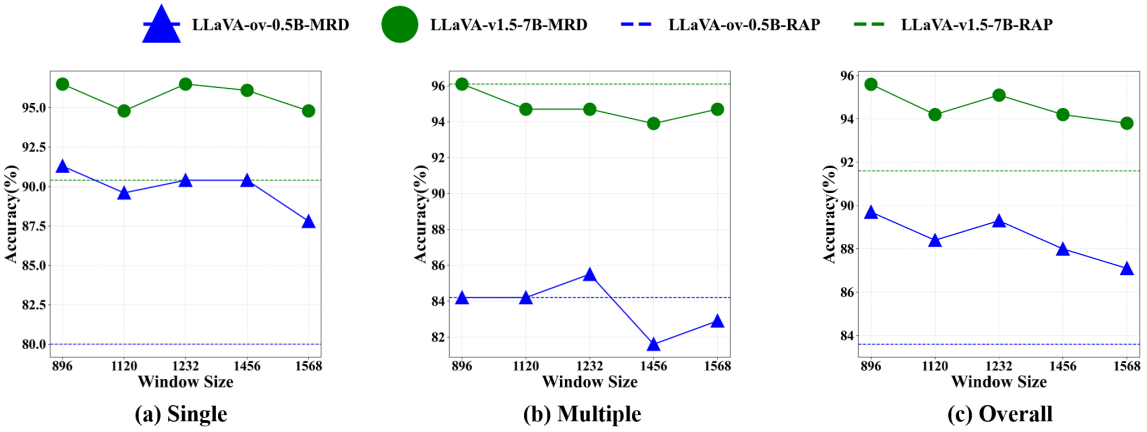


Figure 9. The effect of the detection window size in ***MRD***.

the query—exhibits undesirably high similarity scores.

The proposed ***MRD*** addresses these issues through two key mechanisms. First, the multi-resolution semantic fusion module mitigates semantic inconsistencies caused by crop partitioning by aggregating information across multiple resolutions, which helps preserve the integrity of the target object. Second, by incorporating an object detection model to explicitly localize candidate regions, ***MRD*** enhances the similarity scores of true target regions while suppressing irrelevant responses. As illustrated in Fig. 11, the resulting semantic–detection fusion maps exhibit clearer contrast between target objects and background regions compared with those produced by *RAP*. Consequently, false positives are significantly reduced, enabling more accurate retrieval of target-related crops during the search process.

C.2. Examples of Multi-object Perception Task

Fig. 12 presents two qualitative examples of the multi-object perception task from each HR benchmark, comparing the performance of ***MRD*** and *RAP* using the LLaVA-v1.5-7B backbone. In the multi-object perception task, the retrieval results reveal that *RAP* often preserves only a subset of the target objects while ignoring others when multiple objects must be localized simultaneously. For example, in the first case from V^* *Bench*, *RAP* completely fails to retrieve the pink umbrella. This issue becomes more evident when there is a large scale discrepancy among the target objects. As shown in the second case of V^* *Bench* and the two examples from *HRBench-8K*, *RAP* tends to retain only the dominant large object while neglecting smaller objects that are also relevant to the query. Similar behavior can also be observed in counting scenarios, such as the two examples

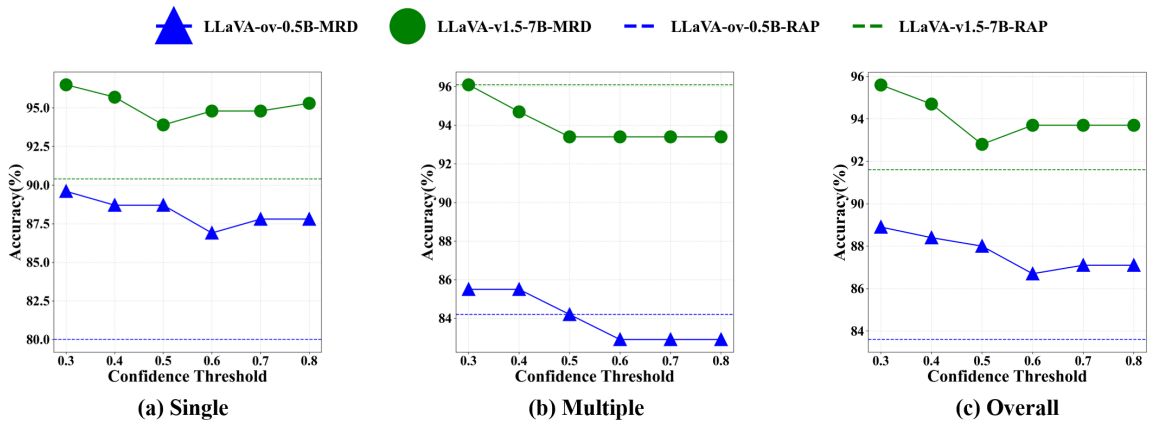


Figure 10. The effect of the detection confidence threshold in ***MRD***.

from *HRBench-4K*, where *RAP* identifies only a subset of the object instances, resulting in incomplete predictions.

These limitations mainly arise from the reliance on semantic similarity maps in *RAP*, where smaller objects or objects with weaker semantic responses may be suppressed during the retrieval process. In contrast, ***MRD*** incorporates an object detection module that explicitly localizes candidate objects before the retrieval stage. This design enables the framework to simultaneously preserve multiple target instances, including small or less prominent objects. As illustrated in Fig. 12, ***MRD*** is able to retain all relevant objects more reliably, leading to more accurate predictions in cross-instance perception tasks.

D. Limitations and Future Works

Despite the substantial improvements achieved by ***MRD*** on single-object perception tasks, the gains on multi-object scenarios are relatively more modest. We believe this limitation mainly comes from the open-vocabulary detection component, which may be less effective in complex scenes containing many small, partially occluded, or densely packed instances. In such cases, the sliding-window detection strategy may exhibit a bias toward salient or dominant objects, thereby reducing recall for less prominent targets.

Another potential limitation is that the current framework relies on fixed window settings and hand-designed fusion hyperparameters. While these settings work well in practice, they may not always be optimal across different image scales, scene complexities, or backbone models. As a result, the performance improvement on more challenging multi-object cases is still constrained by the detection quality and the coverage of candidate regions.

In future work, we plan to investigate more adaptive region exploration strategies, such as content-aware window selection and dynamic multi-scale scanning, to improve small-object recall and better handle crowded scenes. We also aim to explore stronger detection-guided fusion mechanisms that can more effectively balance semantic relevance and spatial localization. Importantly, we hope to preserve the modular and training-free nature of ***MRD***, so that it remains easy to integrate into different multimodal large language models without additional training.

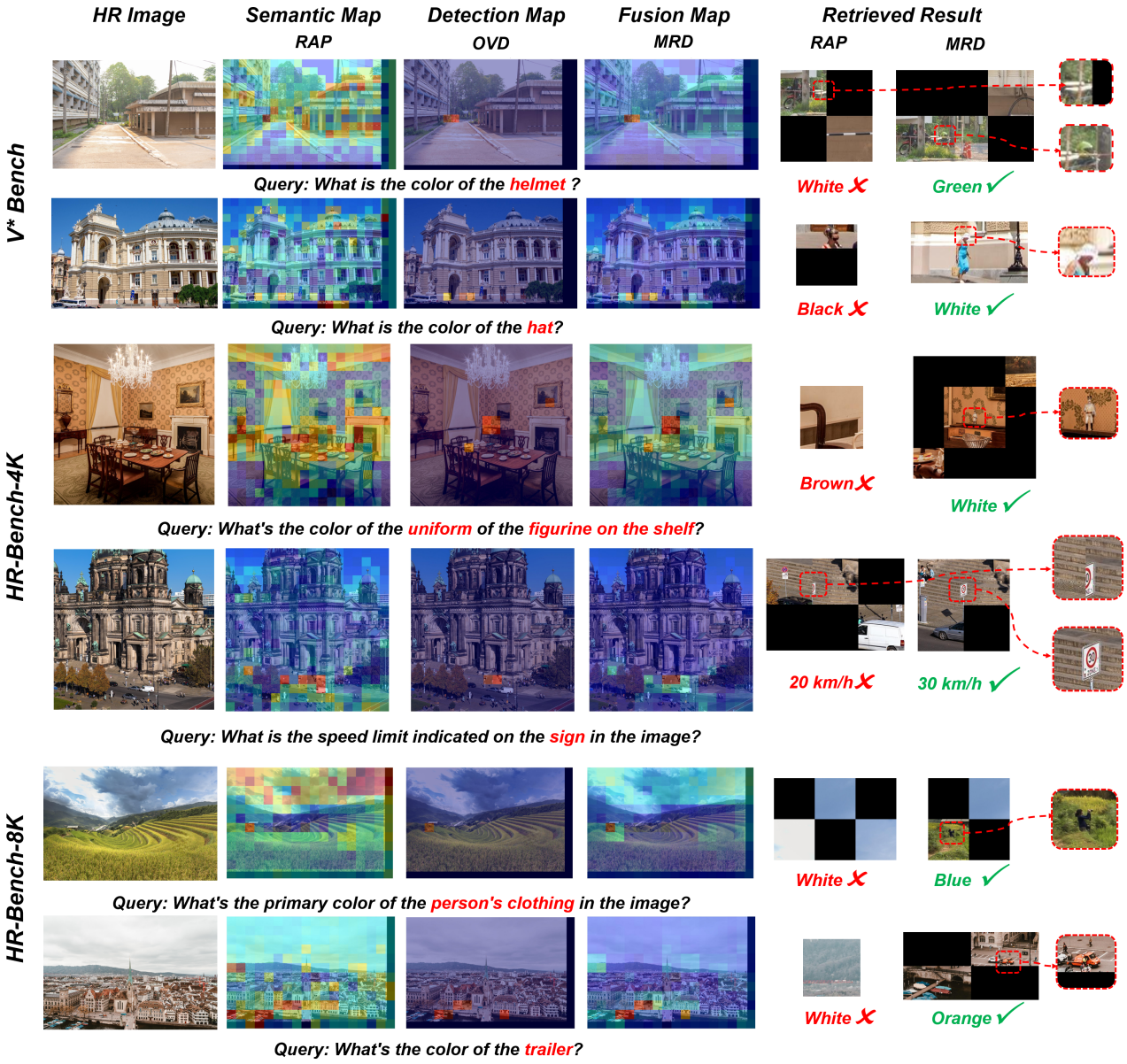


Figure 11. Qualitative examples of single-object perception task. We conduct experiments using LLaVA-v1.5-7B on three HR Benchmarks.

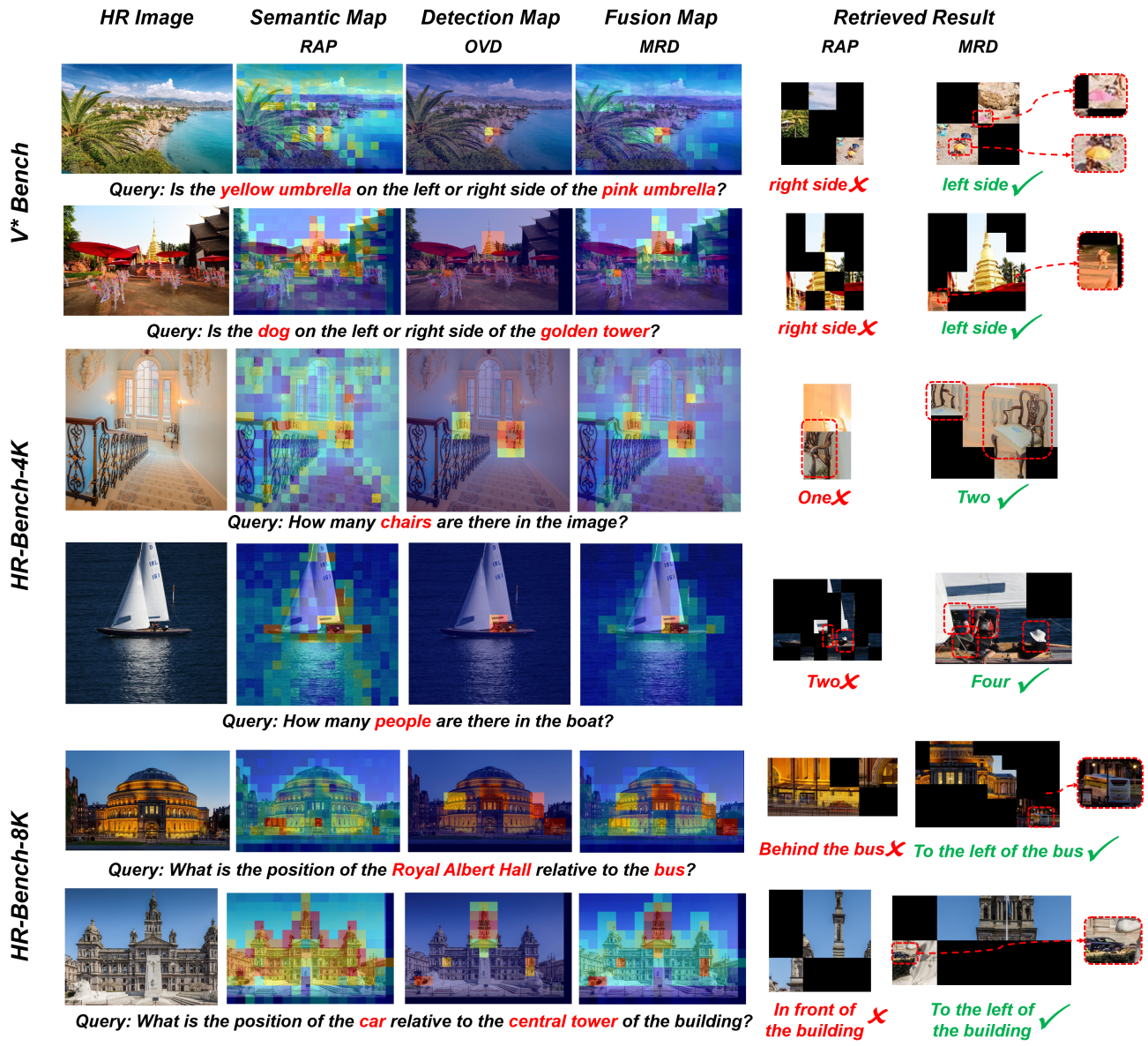


Figure 12. Qualitative examples of multi-object perception task.We conduct experiments using LLaVA-v1.5-7B on three HR Benchmarks.