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

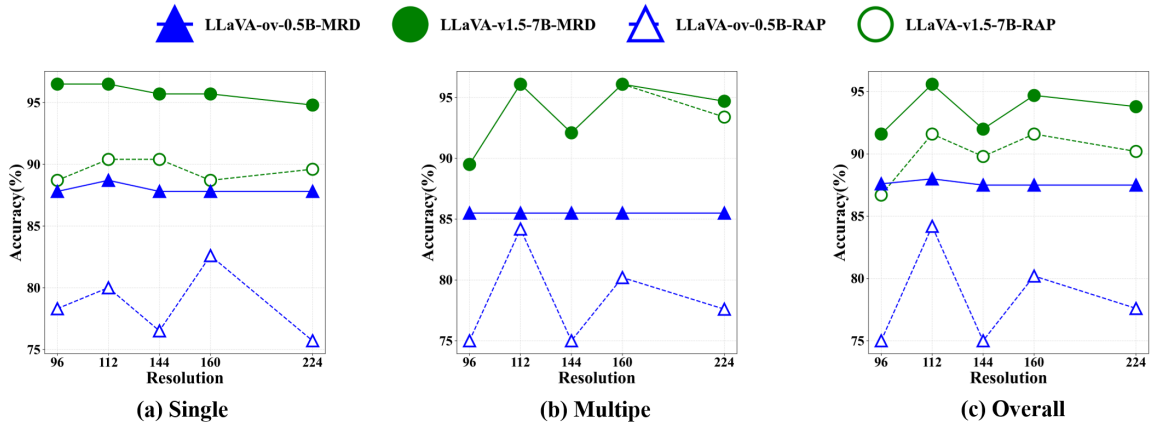

Figure 6. The effect of the resolution of image crops on MLLMs performance. Single and Multiple represent the attribute recognition and spatial reasoning tasks in V^* Bench. (a) Single-object Task. (b) Multi-object Task. (c) Overall Performance.

.

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

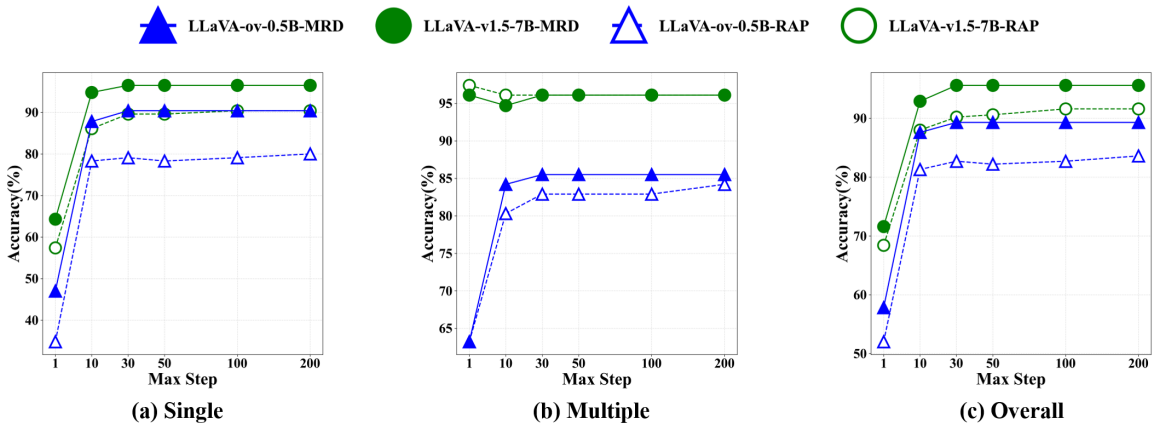

Figure 7. The effect of the maximum search steps in ***MRD*** and *RAP*.

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

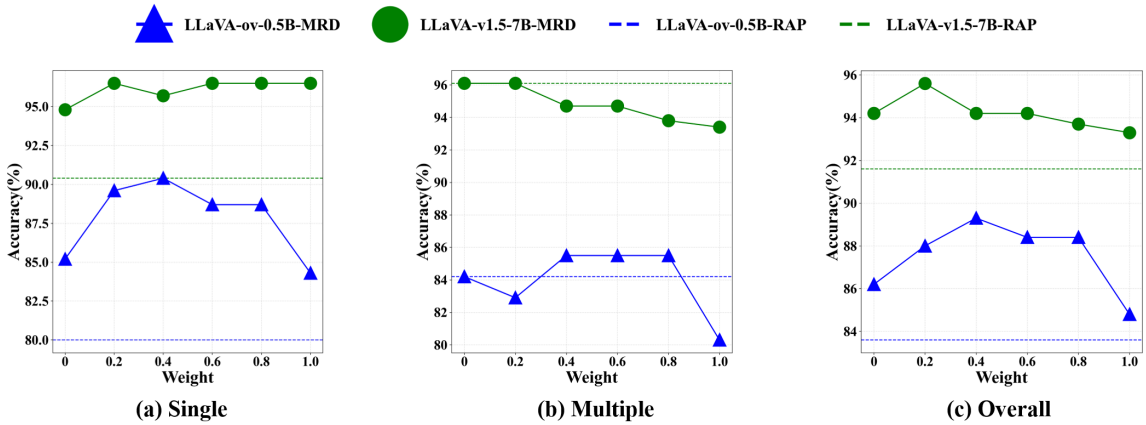

Figure 8. The effect of the detection weight in ***MRD***.

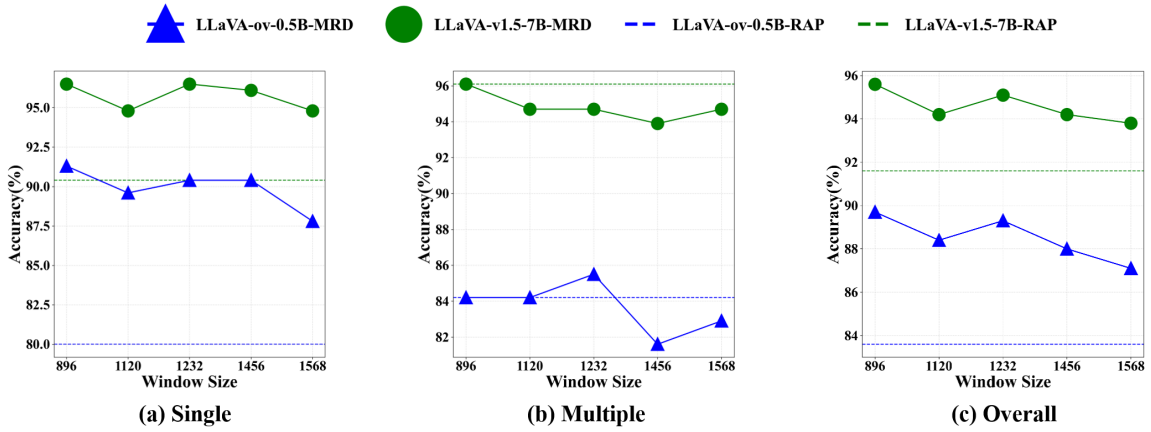

Figure 9. The effect of the detection window size in ***MRD***.

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

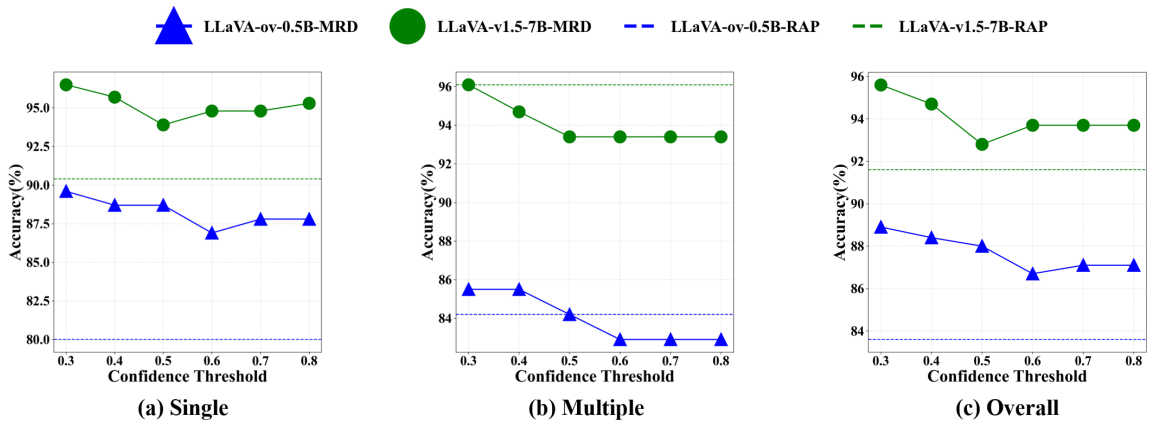

Figure 10. The effect of the detection confidence threshold in ***MRD***.

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.

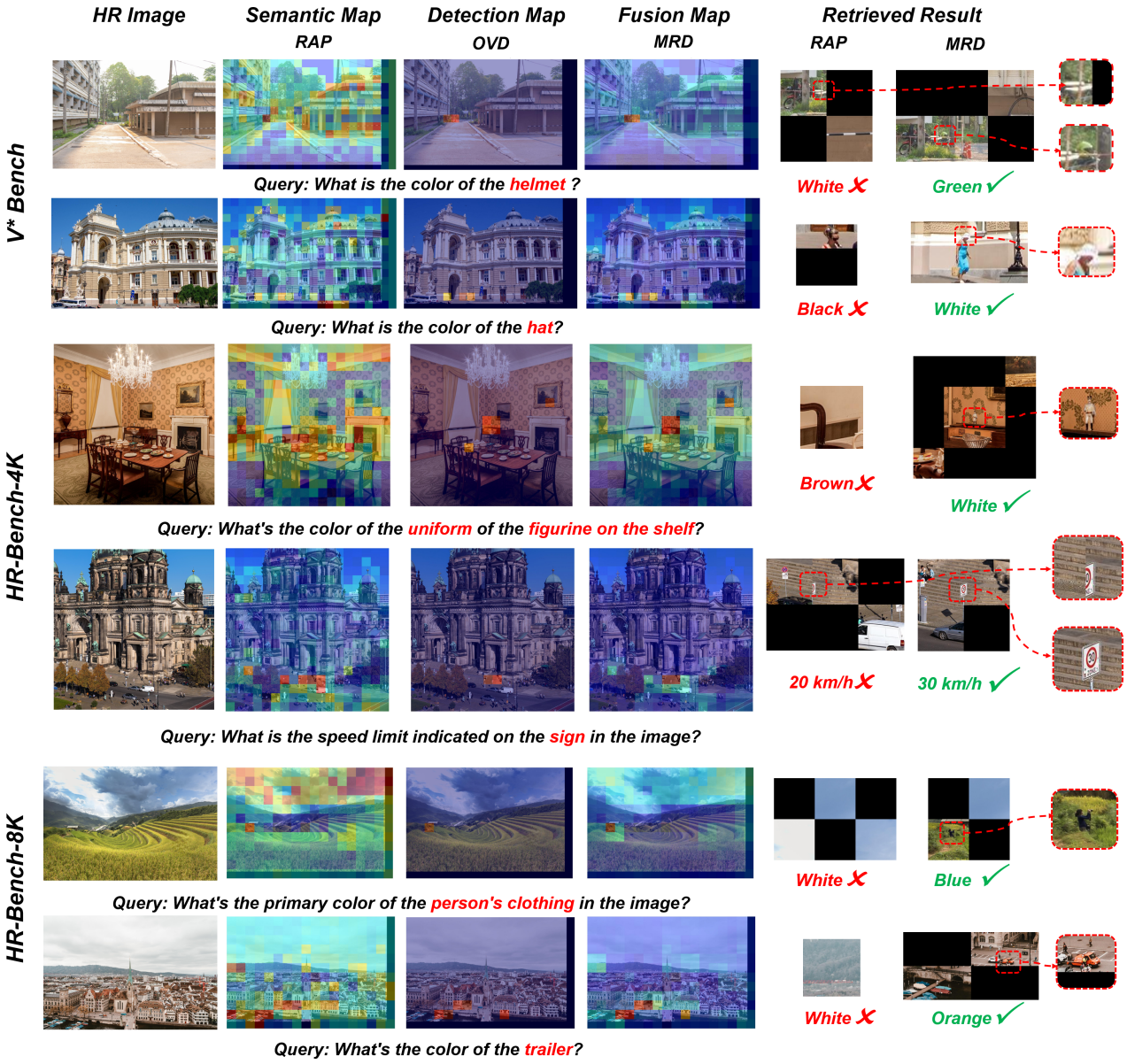

Figure 11. Qualitative examples of single-object perception task. We conduct experiments using LLaVA-v1.5-7B on three HR Benchmarks.

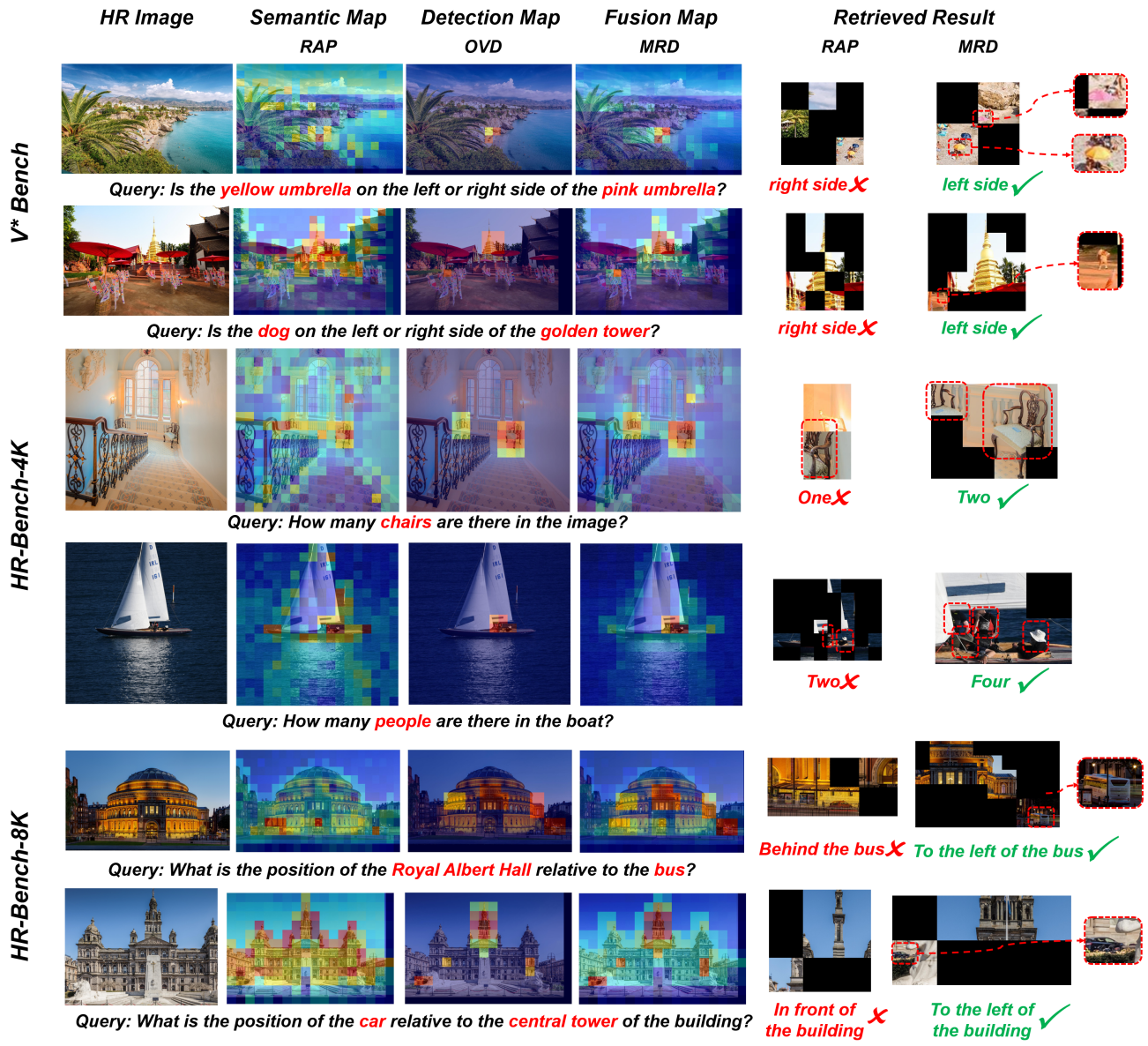

Figure 12. Qualitative examples of multi-object perception task.We conduct experiments using LLaVA-v1.5-7B on three HR Benchmarks.