Towards High-Resolution Visual Perception via Hierarchical Entity Exploration

Ziyu Ma[1⋆], Shidong Yang[1⋆], Yuxiang Ji[1], Yiming Hu[1†], Tongwen Huang[1†], Yong Wang[1‡], Jianfei Cai[2], and Xiangxiang Chu[1]

[1] AMAP, Alibaba Group, China
[2] Monash University, Australia

Abstract. High-resolution (HR) image perception remains a key challenge in multimodal large language models (MLLMs), as fine-grained details are often lost when the image is processed as a whole. Existing methods either require training to teach models where to look or heuristically divide the image into fixed regions, both of which struggle to generalize in complex HR scenes. In this work, we propose Hierarchical Entity Exploration (HEE), a training-free and model-agnostic framework that transforms static image understanding into dynamic, query-guided entity exploration. HEE first evaluates each region using a dual scoring mechanism to determine whether it already contains sufficient evidence to answer the question. If not, it applies object detection within the most promising region to extract fine-grained entities, clusters them into coherent subregions, and organizes them into a multi-level semantic hierarchy for deeper exploration. When deeper regions still fail to yield confident answers, a confidence-guided backtracking mechanism revisits alternative paths to ensure adaptive perception. Extensive results show that HEE outperforms training-free methods like ZoomEye and RAP in both accuracy and efficiency on two complex HR benchmarks (Visual Probe and HR-Bench), across different MLLMs such as Qwen2.5-VL and LLaVA-OneVision. Moreover, HEE demonstrates generalization on the MME-RealWorld benchmark.

Keywords: High-resolution image · MLLMs · Entity exploration

1 Introduction

Current MLLMs [1,3,26,27,33,36,46] operate under a fixed input resolution (e.g., 448×448), a design choice that simplifies computation but compromises the fidelity of high-resolution (HR) images. When HR inputs are uniformly resized, geometric deformation and fine-detail loss inevitably occur, leading to noticeable performance drops in tasks that rely on precise visual cues, such as visual grounding and OCR, where subtle structures are essential [15, 34, 35, 47, 61].

To enhance the model's perception of image details, existing methods often adopt a "locate-then-zoom-in" strategy. Training-based paradigms (e.g., supervised fine-tuning [41] and reinforcement learning [13, 62, 63]) explicitly teach

⋆Equal contribution. †Project lead. ‡Corresponding author.

models to focus on key regions by introducing visual tool-use abilities and corresponding training data. A representative method is DeepEyes [63], which defines a zoom-in tool and optimizes its usage through end-to-end reinforcement learning. During inference, the model actively selects and enlarges task-relevant regions via the visual tool to progressively focus on fine-grained information in high-resolution images. However, such training-based paradigms suffer from critical drawbacks [55, 58], including high costs, lengthy training, and a lack of cross-architecture transferability, which severely limits their practical deployment.

Apart from studies that enhance perception through tool-augmented training, another line of research explores training-free approaches that rely on heuristic exploration to progressively locate task-relevant regions in high-resolution images [21, 42, 59]. Within this paradigm, the core distinction lies in how visual input is effectively explored. ZoomEye [42] adopts uniform spatial partitioning (Fig. 1(a)) and recursively selects sub-regions with higher vision–language relevance for deeper exploration, but such grid-based division often fragments semantically coherent entities and scatters critical visual cues. RAP [49] processes the image differently. Specifically, it employs fixed-scale (e.g., 224/336 pixels) patch partitioning followed by scoring and top-k selection over the entire image (Fig. 1(b)), and then reconstructs query-relevant crops according to their spatial positions into a coherent view for reasoning. Although these methods differ in search strategies, with ZoomEye performing recursive tree search with backtracking and RAP performing one-shot multi-crop retrieval with spatial layout reconstruction, they share the same fundamental limitation: the partitioning boundaries are purely geometric and entirely agnostic to object boundaries. This leads to two problems: (1) *Entity fragmentation*, where objects straddling partition boundaries are split into incomplete pieces, leaving no single candidate that contains the full target information (as also evidenced by the case study in [49]); and (2) *Background interference*, where each geometrically defined region inevitably includes task-irrelevant background content. In both cases, the core operation is to crop a sub-region and resize it as input to the model.

To quantify the impact of background interference, we conduct a hierarchical decoupling analysis (Fig. 1(d)). A natural hypothesis is that cropping works by resizing the target region to a more suitable input resolution, enabling the model to perceive finer details. However, our background masking experiment challenges this assumption: without any cropping, simply removing background pixels progressively yields a monotonic accuracy increase across all models (Fig. 1(d, top); e.g., Qwen2.5-VL-7B: 56%→63%; InternVL2.5-8B: 65%→70%). The zoom-in experiment further corroborates this finding (Fig. 1(d, bottom)): contracting the view toward the GT area improves accuracy, while expanding the canvas with zero-semantic white pixels degrades performance. Together, these results reveal a key insight: the primary gain of cropping comes from background removal, not target refinement. This directly explains the bottleneck of geometry-based partitioning: regardless of how sophisticated the search strategy is, as long as the partitioning is agnostic to object boundaries, the selected regions inevitably con-

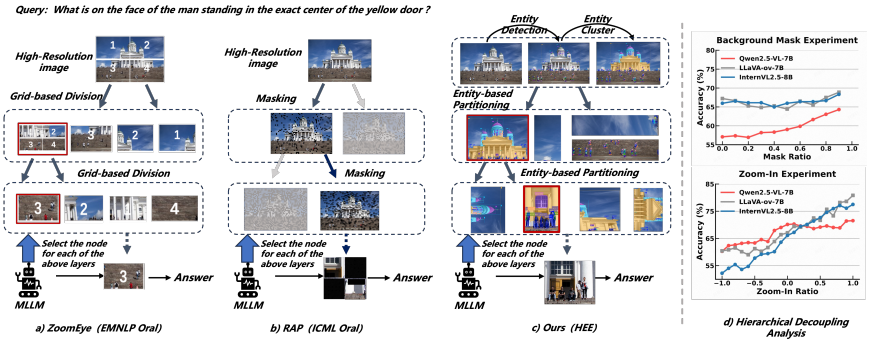

Fig. 1: Illustration of two common approaches (i.e., ZoomEye, and RAP) for handling high-resolution (HR) images, alongside our HEE. The right subfigures show our hierarchical decoupling analysis, including the background masking and zoom-in experiment.

tain background content that harms reasoning. Our conclusions are consistent with [29], further validating the necessity of entity-aware visual perception.

Guided by these findings, we propose Hierarchical Entity Exploration (HEE), a training-free and model-agnostic framework for high-resolution visual perception. Unlike geometry-driven partitioning, HEE adopts entity-driven semantic decomposition (Fig. 1(c)): a frozen object detector discovers object-level entities at each level, so that the search space is organized around complete visual objects rather than being sliced by arbitrary geometric boundaries. This directly addresses both problems: detector-produced bounding boxes are naturally centered on complete objects, avoiding entity fragmentation while achieving a higher foreground ratio than geometric patches. Specifically, HEE first evaluates each visual region via a dual scoring mechanism to determine whether sufficient evidence exists to answer the question. If not, HEE selects the highest-scored region and applies a frozen object detector to extract object-level entities, which are then clustered and merged into semantically coherent sub-nodes forming the next-level search space. This process repeats recursively, with the foreground ratio progressively increasing at each level. When an exploration path fails to yield a confident answer, a backtracking mechanism returns to higher levels to explore new regions, preventing irrecoverable misses due to single-path errors.

HEE achieves significant performance gains across multiple high-resolution benchmarks and model families. On Visual Probe, it improves Qwen2.5-VL-7B from 39.1% to 67.4% on the Easy split and from 23.9% to 50.0% on the Hard split; on HR-Bench, it lifts overall accuracy from 66.5% to 73.9% on 4K and from 62.1% to 71.1% on 8K. On the MME-RealWorld benchmark, HEE brings consistent improvements across five diverse tasks (e.g., +7.5% on Monitoring, +7.2% on Remote Sensing), confirming its robustness in real-world visual conditions. Compared with existing training-free methods such as RAP and ZoomEye, HEE consistently achieves higher accuracy and less computational overhead. These results highlight the potential of HEE as a general, and practical solution for high-resolution visual perception. Our contributions can be summarized as:

- We identify via controlled decoupling experiments that background interference is an important factor limiting geometry-based partitioning. Motivated by this, we propose Hierarchical Entity Exploration (HEE), a training-free and model-agnostic framework that replaces geometric partitioning with entity-driven semantic decomposition.
- HEE introduces entity-based node partitioning, a dual scoring mechanism, and confidence-guided backtracking to enable dynamic, fine-grained visual exploration organized around complete visual objects rather than arbitrary geometric regions.
- HEE consistently outperforms strong baselines and state-of-the-art training-free methods (i.e., RAP, ZoomEye) across multiple open source MLLMs (e.g., Qwen2.5-VL, LLaVA-OneVision) and diverse high-resolution benchmarks (Visual Probe, HR-Bench, and MME-RealWorld).

2 Related Work

Modern MLLMs typically consist of a pretrained visual encoder [5, 10, 14, 17, 28, 38], a frozen language model [2, 9, 43, 44, 53, 56], and a lightweight multimodal connector [8, 22, 23, 37, 39, 45, 50]. To match the resolution used in pretraining (e.g., 336×336 in LLaVA [28]), input images are usually resized, which can blur or distort fine details in HR settings. Existing solutions to this challenge mainly fall into three categories: 1) cropping-based, 2) encoder-based, and 3) search-based methods.

Cropping-Based. These approaches [20, 25, 27] divide HR images into multiple crops, which are encoded independently via ViT [10] and fed into the LLM. While easy to implement, they lack global coherence and incur high inference costs as resolution grows.

Encoder-Based. Another line improves HR perception by upgrading the visual backbone. Works like LLaVA-HR [32], ConvLLaVA [11], HD-SAM [16] and MiniGemini-HD [24] adopt ConvNeXt [30] with higher input sizes or attention refinements. Deepseek-VL [31] and Vary [51] further incorporate segmentation priors such as SAM [18]. These methods improve resolution handling but typically require retraining or architecture changes.

Search-Based Methods. To efficiently process high-resolution images, some methods adopt a top-down search paradigm that iteratively narrows the visual scope based on task relevance. DC^2 [48] maintains a visual memory of objects and locations to retrieve relevant crops and reduce detail loss. ZoomEye [42] applies recursive zoom-in over uniformly partitioned grids to locate informative subregions. SEAL [52] integrates reasoning and retrieval in a unified structure to capture essential visual content. RAP [49] employs fixed-scale patch partitioning followed by scoring and top-k selection, then reconstructs a compact view for inference. Unlike these methods, our approach performs entity-driven, multi-level exploration with relevance scoring and confidence-guided backtracking, enabling more reliable high-resolution perception.

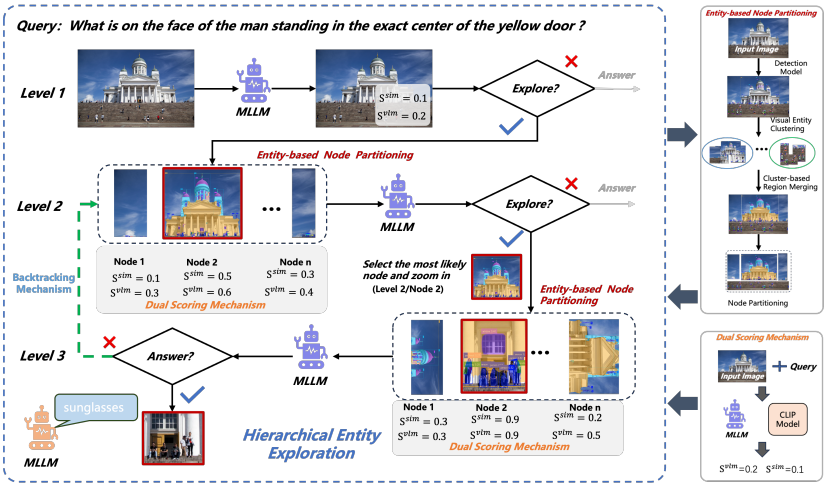

Fig. 2: Overview of the HEE framework. HEE consists of four key components: (1) a dual scoring mechanism that assesses evidence sufficiency by integrating semantic similarity and model confidence; (2) entity-based node partitioning, where detected entities are clustered and merged into semantically coherent sub-regions; (3) a hierarchical exploration strategy that recursively expands the most relevant region; and (4) confidence-guided backtracking, which revisits new nodes when the active branch fails.

3 Method

As illustrated in Fig. 2, Hierarchical Entity Exploration (HEE) is a training-free and model-agnostic framework for high-resolution visual perception. Unlike previous approaches based on uniformly divided regions, HEE formulates perception as an entity-centric hierarchical tree search. At each level, the model first evaluates whether the current visual scope contains sufficient information to answer the question using a dual scoring mechanism that combines semantic similarity and model confidence. If not, the node with the highest score is selected for deeper exploration. For the selected node (or the initial global image), HEE invokes a frozen object detector to extract object-level semantic entities. These entities are then clustered, merged into coherent regions, and partitioned into semantically consistent sub-nodes, which form the next-level search space. This recursive process proceeds adaptively until the model reaches the finest level, where individual entities are directly evaluated. If the current exploration path fails to yield a confident prediction, a backtracking mechanism restores the previous level for alternative exploration, enabling stable, human-like progressive perception without retraining or architectural modification.

3.1 Entity-Based Node Partitioning

After obtaining the selected node or the initial global image, HEE aims to decompose the current visual region into a set of semantically coherent and spatially

complete sub-nodes, thereby constructing the search space for the next level. Given a region $\mathcal{I}^{(l)}$ at level l, a frozen detector $\mathcal{D}$ (implemented with DINO-X [40]) is applied to extract object-level semantic entities:

$$\mathcal{E}^{(l)} = \mathcal{D}(\mathcal{I}^{(l)}) = \{e_i^{(l)}\}_{i=1}^{N_l}, \tag{1}$$

where each entity $e_i^{(l)}$ is represented by its bounding box, category label, and semantic embedding. These detected entities serve as the fundamental visual units for subsequent semantic aggregation and region partitioning. To generate a structured set of candidate nodes for the next-level search, HEE performs K-Means clustering in the semantic embedding space:

$$\mathcal{C}^{(l)} = \text{KMeans}(\mathcal{E}^{(l)}) = \{C_k^{(l)}\}_{k=1}^{K}, \tag{2}$$

where each cluster $C_k^{(l)}$ groups semantically similar or spatially adjacent entities into a coherent conceptual region. For each cluster, the bounding boxes of its member entities are merged into a tight enclosing region:

$$r_k^{(l)} = \text{Merge}(C_k^{(l)}), \tag{3}$$

resulting in a set of explicitly detected semantic areas denoted as $\mathcal{R}_{\text{entity}}^{(l+1)} = \{r_k^{(l)}\}_{k=1}^{K}$. To ensure complete spatial coverage of the scene, HEE also includes the portions of the image that are not covered by any detected entity as independent residual regions:

$$\mathcal{R}_{\text{residual}}^{(l+1)} = \text{MaskDiff}\left(\mathcal{I}^{(l)}, \bigcup_{k=1}^{K} r_k^{(l)}\right), \tag{4}$$

where MaskDiff$(\cdot)$ denotes the pixel-wise difference between the current region mask and the union of all detected entity regions. Finally, the next-level node set is obtained by combining the entity-based and residual regions:

$$\mathcal{R}^{(l+1)} = \mathcal{R}_{\text{entity}}^{(l+1)} \cup \mathcal{R}_{\text{residual}}^{(l+1)}. \tag{5}$$

Each node in $\mathcal{R}^{(l+1)}$ preserves both visual and semantic features and acts as a potential branch in the hierarchical exploration tree. Through this process, HEE maintains semantic consistency while ensuring full spatial coverage, transforming unstructured visual inputs into an organized set of semantic nodes. The resulting node set $\mathcal{R}^{(l+1)}$ provides the structural foundation for the subsequent dual scoring stage, which evaluates task relevance and determines whether further recursive exploration is required.

3.2 Dual Scoring Mechanism

To assess the relevance between the current visual region and the question, HEE adopts a dual scoring mechanism that jointly considers *semantic similarity* and

task relevance. This mechanism combines the semantic alignment capability of a frozen vision-language matching model (e.g., CLIP) with the reasoning confidence of the target VLM, providing a complementary measure of visual-question consistency.

For each candidate region r (a cluster region $C_k^{(l)}$ or an entity $e_i^{(l)}$ at the final level), two scores are computed:

$$s^{\text{sim}}(r) = S(q, r), \quad s^{\text{vlm}}(r) = M(r, q), \tag{6}$$

where $S(q, r)$ denotes the semantic similarity between the question q and region r, and $M(r, q)$ is the VLM's confidence that r provides sufficient evidence to answer q. The final score is the weighted sum of the two:

$$S(r) = \lambda\, s^{\text{vlm}}(r) + (1 - \lambda)\, s^{\text{sim}}(r). \tag{7}$$

The region with the highest score:

$$r^* = \arg\max_r S(r), \tag{8}$$

is selected for further exploration, while a high $S(r)$ value at the current level indicates that the question can already be answered without deeper search. If all regions fall below a confidence threshold τ, HEE initiates backtracking to re-examine alternative candidates from the previous level.

3.3 Adaptive Recursive Exploration

Building upon the dual scoring results, HEE performs adaptive recursive exploration to progressively refine its visual focus. Given the scored region set $\mathcal{R}^{(l+1)}$, the model first determines whether the current level provides sufficient confidence to answer the question. If the highest score $\max_r S(r)$ exceeds the threshold τ, the exploration terminates, and the corresponding region is used for answer generation. Otherwise, HEE selects the most relevant region $r^* = \arg\max_r S(r)$ and applies entity-based partitioning to this region for deeper exploration.

At each level, all candidate regions are scored to preserve a global ranking of task relevance, but only the top-ranked region is expanded into finer sub-nodes, while other regions remain as alternative candidates for potential backtracking. This process continues until one of the following stopping conditions is met: (i) the confidence score surpasses $\tau = 0.5$; (ii) the number of detected entities falls below the minimum threshold η (e.g., only one or two entities remain); or (iii) the maximum recursion depth is reached.

If none of these conditions are met and the current region still lacks sufficient evidence, the system triggers a backtracking step that returns to the previous level and re-evaluates sub-optimal nodes. Through this adaptive search and verification loop, HEE dynamically balances depth and precision. This allows the model to zoom from coarse regions to fine-grained entities while maintaining semantic completeness and stable decision consistency across levels.

3.4 Confidence-Guided Backtracking

When the recursive exploration reaches the final layer and no candidate entity achieves sufficient confidence, HEE activates a confidence-guided backtracking procedure to revisit alternative search paths. This mechanism prevents the model from committing to unreliable local decisions and ensures stable reasoning under uncertain visual conditions.

Specifically, after the final-level scoring, if all entities $\{e_i\}$ yield scores below the confidence threshold τ, i.e.,

$$\max_i S(e_i) < \tau, \tag{9}$$

the system returns to the previous level and re-examines the next-best candidate region (e.g., Top-2, Top-3, etc.) based on the stored global ranking from that level. The selected region is then expanded again using entity-based partitioning and re-evaluated through the dual scoring mechanism. This backtracking process continues iteratively until a valid region surpasses the confidence threshold or all candidates are exhausted. This backtracking mechanism introduces an adaptive verification loop into the hierarchical search, allowing HEE to recover from low-confidence paths and maintain semantic consistency across levels.

4 Experiments

4.1 Evaluation Setup

We evaluate HEE on two high-resolution benchmarks. HR-Bench [48] includes 8K-resolution images for fine-grained perception, covering single-instance (FSP) and cross-instance (FCP) tasks, with a cropped 4K variant also provided. To further assess robustness, we use the Visual Probe Dataset (VisualProbe) [19], a challenging visual search benchmark containing 4,000 training and 500 testing samples. VisualProbe features 4K-resolution images with small targets, dense distractors, and trial-and-error queries, and is divided into *easy*, *medium*, and *hard* subsets to reflect varying levels of semantic complexity and visual ambiguity. To evaluate real-world utility, we include MME-RealWorld [60], a large-scale benchmark with 13,366 high-resolution images and 29,429 human-verified QA pairs across 43 tasks. It poses substantial challenges with expert-annotated questions, diverse domains (e.g., OCR, remote sensing), and image resolutions up to 5000×5000.

HEE is implemented on top of three recent open-source vision-language models: Qwen2.5-VL-7B [4], InternVL2.5-8B [6], and LLaVA-OneVision-7B [20], without any retraining or architectural modifications. These models serve as our base systems to validate the effectiveness and generality of our HEE.

4.2 Implementation Details

All experiments are conducted on an NVIDIA H20 GPU with 80GB memory. We set the number of entity clusters to 4 and fix the confidence threshold τ to 0.5.

Table 1: Comparison of **HEE** applied to several advanced MLLMs against existing methods on Visual Probe, HR-Bench 4K, and HR-Bench 8K.

Method	Visual Probe [19]			HR-Bench 4K [48]			HR-Bench 8K [48]		
	Easy	Medium	Hard	FSP	FCP	Overall	FSP	FCP	Overall
Open-source MLLMs									
LLaVA-v1.6-7B [27]	–	–	–	49.0	46.8	47.9	37.3	44.3	40.8
LLaVA-HR-X-13B [32]	–	–	–	61.3	46.0	53.6	49.5	44.3	46.9
LLaVA-HR-X-7B [32]	–	–	–	57.8	46.3	52.0	42.0	41.3	41.6
Yi-VL-34B [54]	–	–	–	46.0	42.8	44.4	39.5	38.5	39.0
Closed-source MLLMs									
GPT-4o [12]	47.5	11.2	15.4	70.0	48.0	59.0	62.0	49.0	55.5
Qwen-VL-Max [3]	17.7	6.7	2.8	65.0	52.0	58.5	54.0	51.0	52.5
Baseline and HEE									
Qwen2.5-VL-7B [4]	39.1	26.0	23.9	81.5	51.5	66.5	73.3	51.0	62.1
-w/ HEE	**67.4**	**47.0**	**50.0**	**92.3**	**55.5**	**73.9**	**88.5**	**53.8**	**71.1**
$\Delta(\uparrow)$	+28.3	+21.0	+26.1	+10.8	+4.0	+7.4	+15.2	+2.8	+9.0
InternVL2.5-8B [6]	49.6	19.4	17.0	76.8	56.5	66.6	59.5	50.0	54.8
-w/ HEE	**60.3**	**42.2**	**41.5**	**85.0**	**59.8**	**72.4**	**81.8**	**51.3**	**66.5**
$\Delta(\uparrow)$	+10.7	+22.8	+24.5	+8.2	+3.3	+5.8	+22.3	+1.3	+11.7
LLaVA-ov-7B [20]	36.2	12.5	13.4	75.3	54.0	64.6	66.5	48.0	57.3
-w/ HEE	**61.0**	**36.9**	**38.6**	**88.0**	**55.0**	**71.5**	**83.5**	**52.3**	**67.9**
$\Delta(\uparrow)$	+24.8	+24.4	+25.2	+12.7	+1.0	+6.9	+17.0	+4.3	+10.6

Table 2: Performance comparison on the MME-RealWorld [60] across five task categories: OCR, Remote Sensing, Diagram/Table Understanding, Monitoring, and Autonomous Driving. HEE consistently improves Qwen2.5-VL-7B across all dimensions.

Method	OCR	Remote Sensing	Diagram/Table	Monitoring	Autonomous Driving	Avg.
Qwen2.5-VL-7B [4]	84.60	39.22	77.97	38.52	28.83	59.99
-w/ HEE	**87.98**	**46.39**	**79.81**	**46.02**	**31.42**	**63.38**
$\Delta(\uparrow)$	+3.38	+7.17	+1.84	+7.50	+2.59	+3.39

The maximum number of exploration steps $T_{\max}$ is set to 50 unless otherwise noted. We use DINO-X [40] as the default object detector and adopt SigLIP for computing vision-text similarity. To ensure fair comparison with RAP, we use the same textual prompts, and set the model temperature as 0 to eliminate randomness during generation. More details are provided in the Appendix.

4.3 Main Results

Generalization across Model Families. Table 1 presents the high-resolution perception results across three representative MLLM families. On the Visual Probe benchmark, HEE brings consistent improvements across all model families: Qwen2.5-VL-7B gains +28.3%, +21.0%, and +26.1% on the Easy, Medium, and Hard subsets; InternVL2.5-8B gains +10.7%, +22.8%, and +24.5%; LLaVA-OneVision-7B gains +24.8%, +24.4%, and +25.2%. These results show that the entity-centric hierarchical exploration in HEE effectively enhances fine-grained perception across diverse model families by progressively focusing on task-relevant

Table 3: Comparison of RAP, ZoomEye, and HEE on Visual Probe (Easy, Medium, Hard), HR-Bench 4K, and HR-Bench 8K. Qwen refers to the Qwen2.5-VL-7B model. LLaVA refers to the LLaVA-OneVision-7B model.

Model	Method	Visual Probe [19]			HR-Bench 4K [48]			HR-Bench 8K [48]		
		Easy	Medium	Hard	FSP	FCP	Overall	FSP	FCP	Overall
Qwen	RAP [49]	52.5	31.7	34.0	90.8	51.5	71.1	85.5	53.3	69.4
	ZoomEye [42]	55.3	35.5	41.5	87.8	55.0	71.4	83.5	52.3	67.9
	HEE (Ours)	**67.4**	**47.0**	**50.0**	**92.3**	**55.5**	**73.9**	**88.5**	**53.8**	**71.1**
LLaVA	RAP [49]	56.7	32.1	31.1	84.5	45.3	64.9	82.5	38.0	60.3
	ZoomEye [42]	61.0	31.3	31.1	84.5	54.8	69.6	82.3	51.2	66.8
	HEE (Ours)	**61.0**	**36.9**	**38.6**	**88.0**	**55.0**	**71.5**	**83.5**	**52.3**	**67.9**

regions, capturing subtle visual cues in complex high-resolution scenes, and confirming its architecture-agnostic design and generalization capability.

Results on Different Datasets. Table 1 also reports results on two high-resolution benchmarks: HR-Bench and Visual Probe. HR-Bench includes both 4K and 8K variants with fine-grained single-instance (FSP) and cross-instance (FCP) perception. Across both settings, HEE consistently improves all base models, achieving +5.8%–7.4% on 4K and +9.0%–11.7% on 8K (e.g., Qwen2.5-VL-7B gains +7.4% on 4K and +9.0% on 8K). On Visual Probe, a high-resolution benchmark focusing on open-ended VQA, the gains are even larger, showing that HEE is robust across varying resolutions and perception tasks.

Results on Real-World Scenarios. We further evaluate HEE on the MME-RealWorld [60], which covers five real-world categories: OCR, Remote Sensing, Diagram & Table Understanding, Monitoring, and Autonomous Driving. Table 2 reports the results and we can find: (i) In scenarios with high demands for fine-grained perception (e.g., Remote Sensing and Monitoring), HEE brings large gains, demonstrating its ability to capture fine-grained details in complex scenes. (ii) In other categories such as Autonomous Driving, HEE still achieves improvements, showing task generalization across diverse real-world conditions.

Comparison with Existing Methods. We also benchmark HEE against several recent high-resolution test-time methods, including RAP [49], and ZoomEye [42]. As shown in Table 3, HEE consistently achieves the highest overall accuracy across all benchmarks and model families. Compared to RAP and ZoomEye, which rely on geometric partitioning, HEE exhibits more stable cross-resolution performance, highlighting the advantage of its hierarchical, entity-driven exploration.

Table 4: Efficiency comparison of RAP, ZoomEye, and HEE on HR-Bench-4K using Qwen2.5-VL-7B. "Throughout" denotes the number of samples processed per minute, "Time" denotes the overall inference time.

Method	Throughout↑	Time↓	Acc. (%)
RAP [49]	1.6	126min	71.1
ZoomEye [42]	4.3	46min	71.4
HEE (Ours)	**8.3**	**24min**	**73.9**

Furthermore, to validate the practical efficiency of our method, we compare HEE with representative state-of-the-art high-resolution methods on HR-Bench-4K using Qwen2.5-VL-7B under the same experimental setup and hardware environment. As shown in Table 4, HEE achieves the best accuracy (73.9%) while

requiring the least total inference time (24 minutes), outperforming RAP (71.1%, 126 minutes) and ZoomEye (71.4%, 46 minutes). The efficiency gain primarily comes from HEE's hierarchical entity selection, which avoids exhaustive region expansion and focuses computation only on semantically relevant areas. This demonstrates that our structured, entity-centric exploration not only enhances visual precision but also significantly reduces redundant computation, making HEE both effective and practically deployable for high-resolution visual tasks.

4.4 Ablation Studies

Unless otherwise specified, we use Qwen2.5-VL-7B as the default setting for experiments on the Visual Probe, including the Easy, Medium, and Hard subsets.

Effects of different components of HEE. Table 5 reports the contribution of each core component in HEE on the Visual Probe benchmark. Removing the entity exploration module and replacing it with uniform grid partitioning (*w/o Entity Exploration*) leads to performance drops of 5.6%, 2.3%, and 12.3% on the Easy, Medium, and Hard subsets, respectively. This suggests that uniform partitioning fails to capture the semantic organization of the image, while the entity-centric exploration provides a more structured representation that enables the model to focus precisely on task-relevant regions. Removing the confidence-guided backtracking module (*w/o Backtracking*) results in moderate decreases of 7.0%, 2.2%, and 5.2%, indicating that backtracking helps prevent error accumulation and allows recovery from low-confidence exploration paths. Overall, these results demonstrate that entity exploration, and backtracking play complementary roles in structured perception, and dynamic correction, jointly contributing to the robustness of HEE in high-resolution visual perception.

Table 5: Ablation study of HEE on the Visual Probe benchmark. Removing any component leads to a clear drop in accuracy.

Setting	Easy	Medium	Hard
w/o Entity Exploration	61.7	44.7	37.7
w/o Backtracking	60.3	44.8	41.5
HEE (Ours)	**67.3**	**47.0**	**50.0**

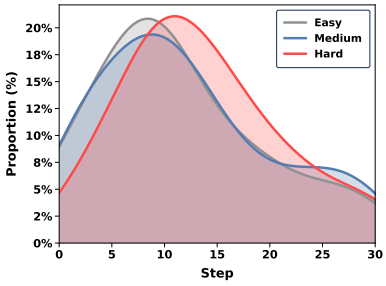

Fig. 3: Distribution of interaction steps in HEE. We plot the step distribution on the Easy, Medium, and Hard splits of Visual Probe using Qwen2.5-VL-7B.

Distribution of Interaction Steps. Fig. 3 reports the distribution of exploration steps required to reach the final decision across the Easy, Medium, and Hard splits of Visual Probe. All three splits exhibit a clear unimodal pattern, with the majority of queries terminating within 8–15 steps. The Easy and Medium splits display nearly identical distributions, both peaking around 10 steps, indicating that HEE identifies the correct entity clusters with minimal search depth. The Hard split shows a slightly right-shifted peak and a heavier tail, reflecting increased ambiguity and smaller

Table 6: Comparison of MLLM performance with different clustered-entity numbers on Visual Probe benchmark.

Model	#E.	Hard	Medium	Easy	Avg. S.
Qwen2.5-VL-7B	–	23.9	26.0	39.1	29.7
w/ HEE	-	40.5	45.9	60.3	48.9
w/ HEE	2	42.5	45.5	61.7	49.9
w/ HEE	4	**50.0**	**47.0**	**67.3**	**54.7**
w/ HEE	8	37.7	45.1	62.4	48.4

Table 7: Ablation study of the scoring formulation in Eq. (7) on Visual Probe using Qwen2.5-VL-7B.

Setting	Hard	Medium	Easy	Average
w/o $S(q,r)$	43.4	45.9	64.5	51.3
w/o $M(r,q)$	21.7	42.9	56.0	40.2
$\lambda = 0.2$	41.5	44.8	65.9	50.7
$\lambda = 0.7$	46.2	46.6	66.7	53.2
$\lambda = 0.5$	**50.0**	**47.0**	**67.3**	**54.7**

visual targets in these examples. Even so, more than 60% of hard cases converge within 15 steps. Overall, these results demonstrate that the hierarchical entity decomposition effectively suppresses irrelevant branches and avoids exhaustive search, enabling HEE to locate task-relevant regions efficiently even in high-resolution, cluttered scenes.

Effect of Clustered-Entity Number. Table 6 analyzes the effect of different clustered-entity numbers on Visual Probe using Qwen2.5-VL-7B. Performance improves steadily as the number of clusters increases from 2 to 4, reaching 67.3%, 47.0%, and 50.0% on the Easy, Medium, and Hard subsets, respectively. This trend shows that a moderate number of entities helps the model capture diverse yet compact semantic regions, facilitating more reliable hierarchical exploration. When the number increases to 8, the performance drops notably, indicating that excessive partitioning produces fragmented representations and redundant search paths that hinder reasoning. These results reveal that the number of entities plays a role in balancing semantic diversity and exploration efficiency.

Table 8: Analysis of detector coverage versus answer correctness. Results are based on Qwen2.5-VL-7B evaluated on Visual Probe.

Detection \ Result	Correct	Wrong
Detected	63.83%	24.82%
Not Detected	3.55%	7.80%

Effect of the Dual Scoring Mechanism. Table 7 studies the impact of the dual scoring function defined in Eq. (7). Removing the semantic similarity term $S(q,r)$ or VLM's confidence term $M(r,q)$ causes a clear performance drop, with the latter showing a stronger effect (40.2% vs. 51.3% on average). This suggests that both components contribute complementary cues: $S(q,r)$ captures semantic relevance between the question and visual region, while $M(r,q)$ reflects model-level confidence in visual–language alignment. Varying the balance factor λ further reveals stable performance within a moderate range (0.2∼0.7), and the best results are achieved at $\lambda = 0.5$. These findings confirm that the proposed dual scoring mechanism effectively integrates semantic and confidence signals, providing a more reliable measure of task relevance during hierarchical exploration. Besides, removing SigLIP $S(q,r)$ and relying solely on the VLM's confidence still yields a

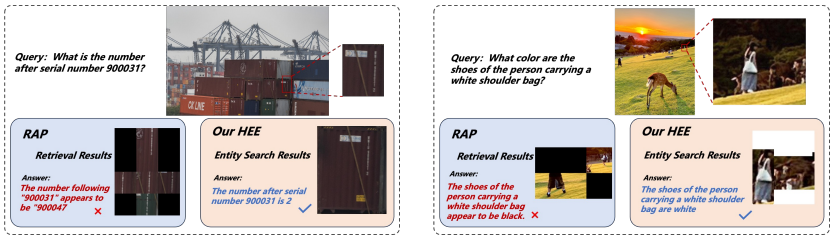

Fig. 4: Comparison with RAP on fine-grained HR visual reasoning. Left: RAP retrieves irrelevant patches and predicts an incorrect serial number; HEE isolates the correct entity. Right: RAP misidentifies shoe color from background-dominant crops; HEE localizes the correct person via the shoulder-bag cue.

51.3 average score, suggesting that region-question relevance is primarily captured by the model itself and its confidence is reliable.

Effect of the Detection Failures. Table 8 categorizes Qwen2.5-VL-7B's predictions on Visual Probe by whether the detector covers the answer-relevant region. Most correct predictions (63.83%) coincide with successful detection, while 3.55% remain correct without it, showing HEE can exploit contextual cues when localization fails. Notably, 24.82% of errors occur despite correct detection, indicating reasoning, not detection, is the dominant bottleneck. Overall, HEE benefits from accurate detection but is not strictly dependent on it.

Effect of Detection and CLIP Encoders. Table 9 examines the influence of different detection models and visual–text alignment encoders on the Visual Probe benchmark. Replacing the detector from DINO-X [40] to YOLO-World [7] leads to a consistent performance drop across all subsets, indicating that accurate entity localization is crucial for reliable hierarchical exploration. Comparing CLIP [38] and SigLIP [57], the latter provides steady improvements under both detectors (e.g., +4.7% on average with DINO-X [40]), suggesting that stronger visual–language alignment benefits the scoring of fine-grained regions. Overall, the results highlight that high-quality detection and semantically aligned CLIP encoders are both essential to maximize HEE's performance in high-resolution visual perception. Furthermore, when we remove SigLIP and replace DINO-X with detection outputs generated directly by the VLM itself, HEE still achieves competitive results (61.0/44.4/45.3 on Easy/Medium/Hard), showing that our HEE remains effective without any external detection or alignment modules.

Table 9: Ablation of different detection models and CLIP encoders on Visual Probe using Qwen2.5-VL-7B.

Det. Model	CLIP Model	Hard	Medium	Easy
DINO-X [40]	CLIP [38]	45.3	45.2	62.4
DINO-X [40]	SigLIP [57]	**50.0**	**47.0**	**67.4**
YOLO-World [7]	CLIP [38]	41.5	45.5	58.9
YOLO-World [7]	SigLIP [57]	41.5	47.0	62.4

Case Study. We present qualitative comparisons in Figs. 4–5. Against RAP, HEE avoids retrieval noise from incomplete or irrelevant patches by progressively narrowing the search through entity-based partitioning, correctly identifying fine-grained targets (e.g., container IDs, shoe colors) that RAP misses.

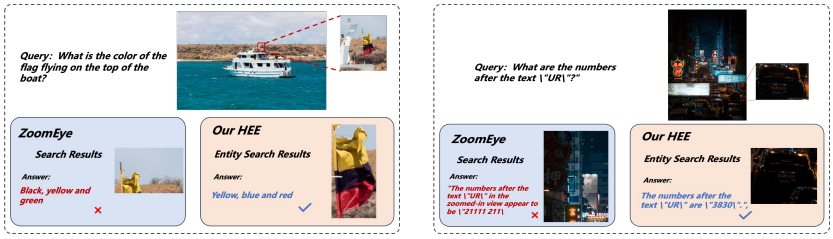

Fig. 5: Comparison with ZoomEye on challenging HR scenarios. Left: ZoomEye zooms into incomplete flag regions and predicts incorrect colors; HEE isolates the full flag entity. Right: ZoomEye is distracted by bright signage and misreads the text; HEE locates the license plate entity and extracts the correct suffix.

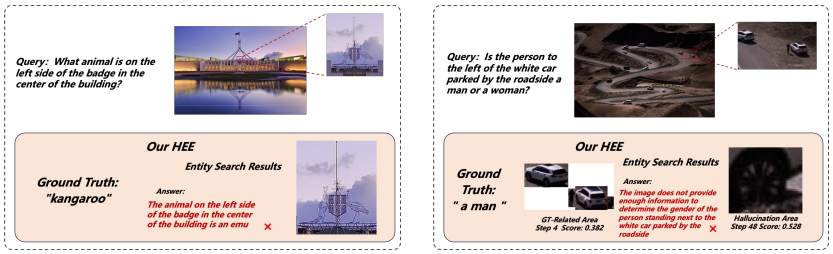

Fig. 6: Failure cases of HEE. Left: HEE correctly locates the emblem region but misclassifies the kangaroo as an emu due to subtle visual similarities. Right: HEE passes through the correct region but drifts to a hallucinated high-confidence area, producing an incorrect answer.

Against ZoomEye, HEE isolates complete semantic units (e.g., flags, license plates) rather than zooming into fragmented regions dominated by background, yielding accurate answers in cluttered scenes. We also analyze two failure modes (Fig. 6): (i) correct localization but incorrect fine-grained recognition, where subtle visual differences (e.g., metallic sculptures) remain ambiguous even after precise decomposition, and (ii) confidence misalignment, where the model visits the correct region but drifts toward a hallucinated high-confidence area.

5 Conclusion

We propose Hierarchical Entity Exploration (HEE), a training-free and model-agnostic framework for high-resolution visual perception. Unlike previous methods based on geometric partitioning, HEE performs entity-centered dynamic exploration through recursive selection, dual scoring, and confidence-guided backtracking. Experiments on Visual Probe, HR-Bench, and MME-RealWorld show consistent and significant gains across diverse multimodal model families. Ablation studies confirm that entity exploration, dual scoring, and backtracking are essential for achieving fine-grained perception and stable reasoning. Moreover, HEE achieves these improvements with low computational overhead, demonstrating an efficient solution for high-resolution scenarios.

References

1. Abdin, M., Jacobs, S.A., Awan, A.A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al.: Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 (2024)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021)
6. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
7. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: CVPR. pp. 16901–16911 (2024)
8. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: CVPR. pp. 2818–2829 (2023)
9. Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., et al.: Qwen2-audio technical report. arXiv preprint arXiv:2407.10759 (2024)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
11. Ge, C., Cheng, S., Wang, Z., Yuan, J., Gao, Y., Song, J., Song, S., Huang, G., Zheng, B.: Convllava: Hierarchical backbones as visual encoder for large multimodal models. arXiv preprint arXiv:2405.15738 (2024)
12. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
13. Ji, Y., Ma, Z., Wang, Y., Chen, G., Chu, X., Wu, L.: Tree search for LLM agent reinforcement learning. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=ZpQwAFhU13
14. Jing, Y., Yang, Y., Wang, X., Song, M., Tao, D.: Meta-aggregator: Learning to aggregate for 1-bit graph neural networks. In: ICCV. pp. 5301–5310 (2021)
15. Jing, Y., Yuan, C., Ju, L., Yang, Y., Wang, X., Tao, D.: Deep graph reprogramming. In: CVPR. pp. 24345–24354 (2023)
16. Ke, L., Ye, M., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F., et al.: Segment anything in high quality. NeurIPS **36**, 29914–29934 (2023)
17. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM Computing Surveys (CSUR) **54**(10s), 1–41 (2022)

18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
19. Lai, X., Li, J., Li, W., Liu, T., Li, T., Zhao, H.: Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. In: ICLR (2026)
20. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. TMLR **2025** (2025)
21. Li, G., Xu, J., Zhao, Y., Peng, Y.: Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In: CVPR. pp. 9098–9108 (2025)
22. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML. pp. 19730–19742 (2023)
23. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. NeurIPS **34**, 9694–9705 (2021)
24. Li, Y., Zhang, Y., Wang, C., Zhong, Z., Chen, Y., Chu, R., Liu, S., Jia, J.: Mini-gemini: Mining the potential of multi-modality vision language models. IEEE Transactions on Pattern Analysis and Machine Intelligence **48**(3), 3530–3543 (2026)
25. Liu, H., You, Q., Han, X., Wang, Y., Zhai, B., Liu, Y., Tao, Y., Huang, H., He, R., Yang, H.: Infimm-hd: A leap forward in high-resolution multimodal understanding. arXiv preprint arXiv:2403.01487 (2024)
26. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR. pp. 26296–26306 (2024)
27. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen: Llava-next: Improved reasoning, ocr, and world knowledge (2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
28. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS **36**, 34892–34916 (2023)
29. Liu, X., Hu, Y., Zou, Y., Wu, L., Xu, J., Zheng, B.: Hide: Rethinking the zoom-in method in high resolution mllms via hierarchical decoupling. In: ICML (2026)
30. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: CVPR. pp. 11976–11986 (2022)
31. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al.: Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525 (2024)
32. Luo, G., Zhou, Y., Zhang, Y., Zheng, X., Sun, X., Ji, R.: Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. In: ICLR (2025)
33. Ma, Z., Gou, C., Hu, Y., Wang, Y., Zhuang, B., Cai, J.: Where and what matters: Sensitivity-aware task vectors for many-shot multimodal in-context learning. Proceedings of the AAAI Conference on Artificial Intelligence **40**(10), 7892–7900 (Mar 2026). https://doi.org/10.1609/aaai.v40i10.37733, https://ojs.aaai.org/index.php/AAAI/article/view/37733
34. Ma, Z., Gou, C., Shi, H., Sun, B., Li, S., Rezatofighi, H., Cai, J.: Drvideo: Document retrieval based long video understanding. arXiv preprint arXiv:2406.12846 (2024)
35. Ma, Z., Li, S., Sun, B., Cai, J., Long, Z., Ma, F.: Gerea: Question-aware prompt captions for knowledge-based visual question answering. arXiv preprint arXiv:2402.02503 (2024)

36. Ma, Z., Yang, S., Ji, Y., Wang, X., Wang, Y., Hu, Y., Huang, T., Chu, X.: Skillclaw: Let skills evolve collectively with agentic evolver. arXiv preprint arXiv:2604.08377 (2026)
37. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. TMLR (2024)
38. 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: ICML. pp. 8748–8763 (2021)
39. Rao, J., Wang, F., Ding, L., Qi, S., Zhan, Y., Liu, W., Tao, D.: Where does the performance improvement come from? -a reproducibility concern about image-text retrieval. In: SIGIR. pp. 2727–2737 (2022)
40. 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)
41. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. NeurIPS **37**, 8612–8642 (2024)
42. Shen, H., Zhao, K., Zhao, T., Xu, R., Zhang, Z., Zhu, M., Yin, J.: Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In: EMNLP. pp. 6613–6629 (2025)
43. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
44. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
45. Wang, F., Ding, L., Rao, J., Liu, Y., Shen, L., Ding, C.: Can linguistic knowledge improve multimodal alignment in vision-language pretraining? ACM Transactions on Multimedia Computing, Communications and Applications **20**(12), 1–22 (2024)
46. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., XiXuan, S., et al.: Cogvlm: Visual expert for pretrained language models. NeurIPS **37**, 121475–121499 (2024)
47. Wang, W., Ding, L., Shen, L., Luo, Y., Hu, H., Tao, D.: Wisdom: Improving multimodal sentiment analysis by fusing contextual world knowledge. In: ACM MM. p. 2282–2291 (2024)
48. Wang, W., Ding, L., Zeng, M., Zhou, X., Shen, L., Luo, Y., Yu, W., Tao, D.: Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In: AAAI. vol. 39, pp. 7907–7915 (2025)
49. Wang, W., Jing, Y., Ding, L., Wang, Y., Shen, L., Luo, Y., Du, B., Tao, D.: Retrieval-augmented perception: High-resolution image perception meets visual rag. In: ICML. vol. 267, pp. 63290–63307 (2025)
50. Wang, X., Li, X., Ding, L., Zhao, S., Biemann, C.: Using self-supervised dual constraint contrastive learning for cross-modal retrieval. In: ECAI. pp. 2552–2559 (2023)
51. Wei, H., Kong, L., Chen, J., Zhao, L., Ge, Z., Yang, J., Sun, J., Han, C., Zhang, X.: Vary: Scaling up the vision vocabulary for large vision-language model. In: ECCV. pp. 408–424 (2024)
52. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms. In: CVPR. pp. 13084–13094 (2024)

53. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
54. Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., et al.: Yi: Open foundation models by 01. ai. arXiv preprint arXiv:2403.04652 (2024)
55. Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Yue, Y., Song, S., Huang, G.: Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? In: NeurIPS. vol. 38, pp. 57654–57689 (2025)
56. Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al.: Glm-130b: An open bilingual pre-trained model. In: ICLR (2023)
57. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV. pp. 11975–11986 (2023)
58. Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y.J., Ma, Y.: Investigating the catastrophic forgetting in multimodal large language model. In: CPAL (2023)
59. Zhang, J., Khayatkhoei, M., Chhikara, P., Ilievski, F.: Mllms know where to look: Training-free perception of small visual details with multimodal llms. In: ICLR (2025)
60. Zhang, Y.F., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., et al.: Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In: ICLR (2025)
61. Zhang, Y., Liu, Y., Guo, Z., Zhang, Y., Yang, X., Zhang, X., Chen, C., Song, J., Yao, Y., Chua, T.S., Sun, M.: Llava-uhd v2: Exploiting hierarchical vision granularity in mllms via inverse semantic pyramid. AAAI **40**(15), 12934–12942 (2026)
62. Zhao, K., Zhu, B., Sun, Q., Zhang, H.: Unsupervised visual chain-of-thought reasoning via preference optimization. In: ICCV. pp. 2303–2312 (2025)
63. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In: ICLR (2026)