This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DyFo: A Training-Free Dynamic Focus Visual Search for Enhancing LMMs in Fine-Grained Visual Understanding

Geng Li[1] Jinglin Xu[2] Yunzhen Zhao[3] Yuxin Peng[1] *
[1]Wangxuan Institute of Computer Technology, Peking University
[2]School of Intelligence Science and Technology, University of Science and Technology Beijing
[3]Tencent Beijing Research, Beijing, 100193, China
ligeng@stu.pku.edu.cn, xujinglinlove@gmail.com, yunzhenzhao@tencent.com, pengyuxin@pku.edu.cn

Abstract

Humans can effortlessly locate desired objects in cluttered environments, relying on a cognitive mechanism known as visual search to efficiently filter out irrelevant information and focus on task-related regions. Inspired by this process, we propose `DyFo` *(*`Dynamic Focus`*), a training-free dynamic focusing visual search method that enhances fine-grained visual understanding in large multimodal models (LMMs). Unlike existing approaches which require additional modules or data collection,* `DyFo` *leverages a bidirectional interaction between LMMs and visual experts, using a Monte Carlo Tree Search (MCTS) algorithm to simulate human-like focus adjustments. This enables LMMs to focus on key visual regions while filtering out irrelevant content, without introducing additional training caused by vocabulary expansion or the integration of specialized localization modules. Experimental results demonstrate that* `DyFo` *significantly improves fine-grained visual understanding and reduces hallucination issues in LMMs, achieving superior performance across both fixed and dynamic resolution models.*

1. Introduction

Humans can quickly locate desired objects in a cluttered environment or recognize familiar faces in a crowd. This efficient visual information retrieval ability is attributed to a visual processing mechanism known as **visual search**, which has been extensively studied in cognitive and vision sciences [29, 42, 46, 48, 49]. In visual search, dynamic focusing adjustments are taken to efficiently filter out irrelevant visual information and concentrate on task-related regions [34, 36]. For example, when reading, people focus on semantically relevant text regions (fixations) and quickly scan past irrelevant areas (saccades) [13, 31]. This inspires us to recognize the importance of selecting visual inputs based on semantic relevance for efficient visual processing.

As a promising paradigm for achieving general multimodal perception and reasoning capabilities, LMMs have achieved notable performance improvements across a range of multimodal tasks [1, 5, 19, 24, 64]. Some LMMs, such as BLIP-2[19] and LLaVA[24], use fixed-resolution encoders to process visual information. These encoders, like ViT[7], operate at a specified input resolution (224 or 336), converting scaled images into a fixed number of tokens. This approach often faces difficulties with high-resolution image tasks because downsizing images and limiting the number of tokens cause the loss of important details, such as small objects, which can eventually lead to hallucinations[9, 44]. To handle high-resolution images, some works (e.g., Qwen2-VL [45] and LLaVA-Next [17]) divide images into smaller sections. Each section is processed separately to generate fixed number of tokens, which are then combined. This approach supports dynamic resolutions and variable token counts, greatly enhancing LMMs' ability to process high-resolution images. However, real-world visual inputs often consist of complex scenes with many unrelated objects, such as busy streets crowded with pedestrians and vehicles or supermarket shelves filled with products. In fine-grained visual understanding tasks that require focusing on specific visual subregions to answer detailed questions, unrestrictedly increasing visual input may introduce a large amount of irrelevant content, causing interference. Some studies [43, 58] have shown that as irrelevant input increases, hallucinations in large multimodal models (LMMs) become more pronounced. Therefore, merely expanding resolution and input capacity may further introduce hallucinations, ultimately limiting the performance of fine-grained visual understanding tasks.

To enhance LMMs' fine-grained visual understanding [11, 56, 61], some studies [3, 30, 57] use region bounding boxes or masks during training to enable local re-

*Corresponding author.

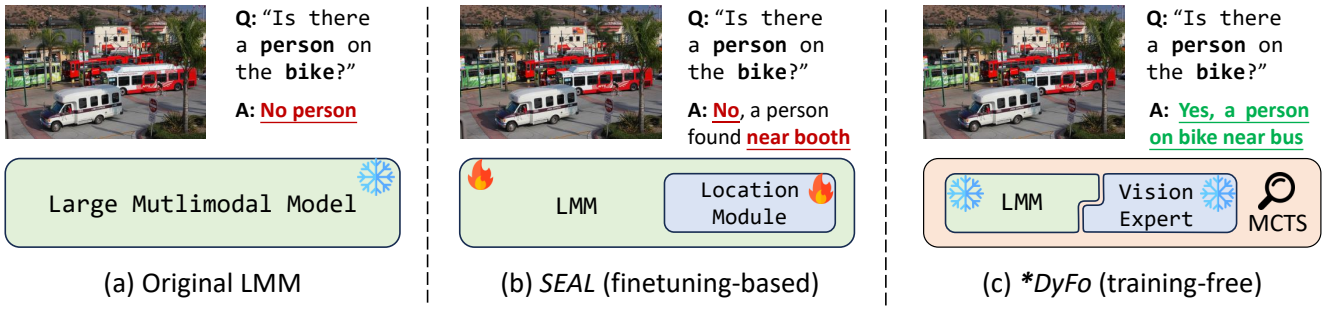


Figure 1. An illustration of three mechanisms for large multimodal models (LMMs) in fine-grained visual understanding tasks.

gion understanding. However, this approach requires user-provided spatial prompts (e.g., boxes or masks) at inference, which are unavailable in real-world applications. SEAL [51] combines visual search with LMMs to assist in automatically locating specific visual regions to improve fine-grained understanding. To support multi-turn search, SEAL utilizes a visual working memory to store intermediate results and fine-tune the system to adapt. Besides, it requires modifications to the LMM's vocabulary and additional localization modules, such as object detection and heatmap generation, to enable visual search. While SEAL achieves strong results, the need for extensive data collection and the adaptation of additional modules limits its applicability to general-purpose LMMs like Qwen2-VL. Moreover, recent study [3, 33, 57, 62] show that methods incorporating extra modules with LMM to support ground task still lag behind visual expert models (e.g., Grounding Dino [25], Lang SAM [26]) in visual task performance, hindering further improvements in visual search.

Inspired by human dynamic focusing adjustments, we propose DyFo (Dynamic Focus), a training-free visual search method that enables LMMs to incorporate the abilities of visual experts, selectively amplifying critical visual information and filtering irrelevant input to enhance fine-grained visual understanding tasks. DyFo leverages bidirectional interaction between LMMs and visual experts to directly focus on key visual regions, and through Monte Carlo Tree Search (MCTS) [40], simulates human-like focus adjustments in visual search. Unlike existing methods like SEAL, DyFo does not require vocabulary expansion or additional dependencies on instruction-tuned localization or heatmap modules. With a consensus-based reward mechanism between the LMM and the visual expert and an action space that simulates human focusing adjustments, DyFo provides an "out-of-the-box" dynamic focusing visual search solution. Experimental results demonstrate that our method effectively enhances fine-grained visual understanding and broadly reduces visual hallucination issues. In summary, our contributions are threefold:

- We propose DyFo, a training-free visual search method that significantly improves fine-grained visual understanding tasks for both fixed and dynamic resolution-supported models.
- We introduce an MCTS-based collaboration paradigm, allowing the use of high-performance visual experts without costly LMM training or architectural changes.
- We validate the superior performance of our method in general hallucination and fine-grained visual understanding tasks, demonstrating its potential to enhance visual capabilities.

2. Related work

Large Multimodal Models. Since the advent of large language models (LLMs), their success in diverse language applications has led to the development of LMMs that integrate vision and language. Initially, LMMs functioned as dispatchers connecting language models with vision experts, as seen in Visual ChatGPT [50], HuggingGPT [39], and MM-REACT [54]. More recently, LMMs have focused on aligning visual and language modalities through training on image-caption pairs or visual question-answering datasets. Methods like LLaVA [24] map image tokens to pre-trained LLM representations via a learned projector, while BLIP-2 [18, 19] employs a query transformer to generate image embeddings from learned queries after extracting image features. MoVA [65] uses an adaptive router to integrate task-specific vision experts with a coarse-to-fine approach. Qwen2-VL [45] enhances high-resolution image understanding through dynamic resolution handling and flexible tokenization.

Visual Search. Inspired by human visual search and its connection to eye movements, early methods simulated this process using Bayesian frameworks with saliency maps [38, 42], deep networks for similarity mapping [60], or inverse reinforcement learning (IRL) to learn search strategies [53]. However, these approaches mainly replicate gaze trajectories without precise target localization and rely on fixed-size attention windows, limiting their flexibility in complex scenes. SEAL [51] integrates localization modules and visual memory to combine visual search with large multimodal models (LMMs) to guide attention regions, but

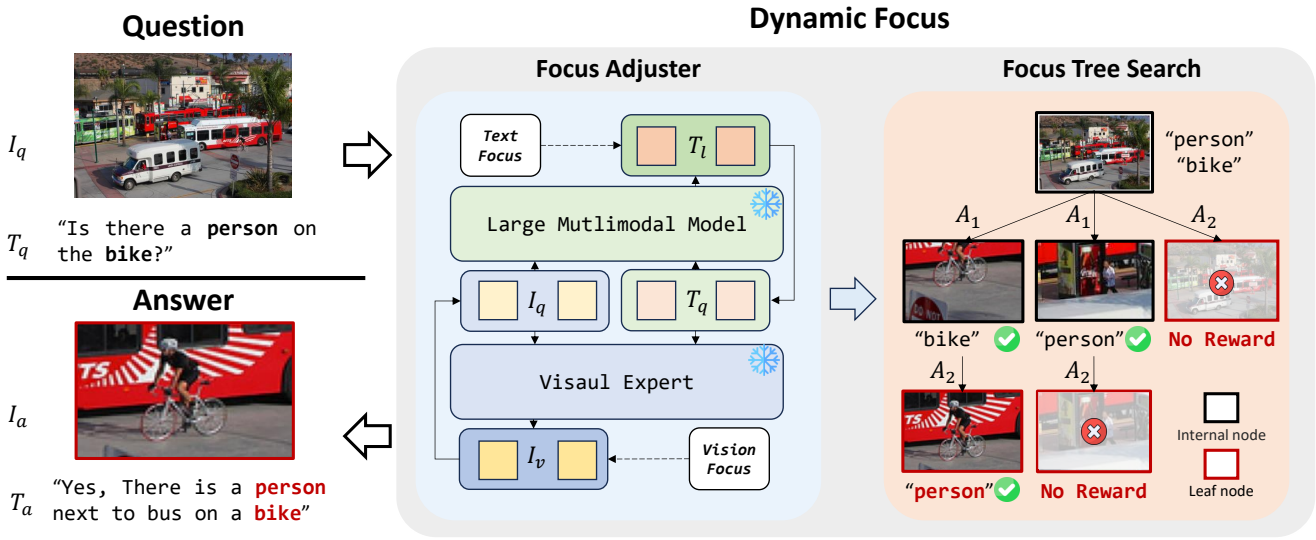


Figure 2. An illustration of DyFo framework, composed by **Focus Adjuster** and **Focus Tree Search** (Section 3.2).

it requires additional training for effective visual search. In contrast, our approach leverages the collaboration between LMMs and visual experts to achieve a dynamic focus mechanism without requiring extra training or module dependencies, providing a "plug-and-play" solution to enhance fine-grained visual understanding.

Visual Experts. Visual experts are models specifically designed to handle vision tasks with both text and image inputs, capable of generating precise image outputs guided by text. Recent studies have leveraged the "Segment Anything Model" (SAM) for text-prompted segmentation [15, 21, 35, 52, 63]. For instance, Grounded-SAM combines text-prompted boxes generated by Grounding DINO with SAM's segmentation function to create a two-stage framework [25, 35]. Fast-SAM matches text and image regions of interest based on CLIP feature similarity [32, 63]. RefSAM uses a cross-modal MLP to project text embeddings into SAM's sparse and dense embeddings, while LISA leverages large multimodal models like LLaVA to extract multimodal embeddings for SAM [21, 24].

3. Method

3.1. Visual Search

In our setup, a large multimodal model (LMM) with parameters θ takes an image $I \in \mathbb{R}^{H\times W\times 3}$ and a text $X = [x_1, ..., x_n]$ as input, where the text contains n tokens, and outputs a text $Y = [y_1, ..., y_m]$ with m tokens. When generating the output text Y, the model sample autoregressively from a probability distribution conditioned on the input I and X. At time step t, the probability of token y_t is

$$y_t \sim p_\theta(y_t|I, X, y_{<t}) \propto \exp \text{logit}_\theta(y_t|I, X, y_{<t}), \quad (1)$$

where logit_θ is the unnormalized log probability of the token y_t, i.e., the value before softmax. The goal of visual search is to find the corresponding image regions $I \in \mathbb{R}^{H'\times W'\times 3}$, where $H' < H$ and $W' < W$, making $\text{logit}_\theta(\hat{y}_t|I, X, y_{<t})$ as high as possible. The $\hat{y}_t$ is the token of the right answer. Recent studies [51] have introduced localization modules that output heatmaps and object boxes to highlight specific areas to enhance the visual grounding ability of models. However, this visual search method relies on the specific architecture of LMMs with extra localization module and introduces substantial costs in grounding data collection and LMM finetuning, hindering its transferability to other models such as Qwen2-VL. We explore a more general visual search strategy by leveraging the collaboration between existing visual experts and LMMs to achieve general and more efficacious visual search.

3.2. Dynamic Focused Visual Search

We propose a training-free visual search method called Dynamic Focus (DyFo), designed to help large multimodal models (LMMs) focus on specific regions of an image through collaboration with visual experts. Specifically, we employ a Monte Carlo Tree Search (MCTS) algorithm, treating focus regions and text as nodes to simulate a human-like, focus-based search process. During this process, the LMM's textual output guides the visual expert to retrieve relevant image regions, while the visual expert's image output, in turn, directs the LMM to adjust its textual focus, as illustrated in Fig. 2.

3.2.1 Focus Adjuster

Although LMMs possess multimodal integration capabilities, they are often limited in fine-grained visual tasks due

to constraints in resolution and hallucination caused by irrelevant content. Whether it's fixed-resolution models (e.g., LLaVA) or dynamic-resolution models (e.g., Qwen2-VL), they struggle to accurately capture local details in complex images like Fig. 1 (a). In contrast, visual expert models excel at precise feature extraction and small object detection but lack comprehensive semantic understanding and robust multimodal integration.

To address these limitations, we propose a Focus Adjuster, a collaborative mechanism guided by action instructions that combines the strengths of both LMMs and visual experts. By achieving mutual enhancement from both textual and visual perspectives, the Focus Adjuster fuses the multimodal understanding of LMMs with the precision of visual experts, significantly improving fine-grained visual search capabilities. As shown in Fig. 3, this integration is realized in the visual search process.

We define the *focus* as $f = (I, T)$, where $I \in I_o$ is the relevant image region selected from the original image I_o, and T represents the primary semantic cue associated with I. The iterative update of focus f proceeds as follows:

$$\begin{cases} T^{i+1} = L(f^i, A^i), \\ I^{i+1} = E(T^{i+1}, I^i, A^i), \\ f^{i+1} = (I^{i+1}, T^{i+1}), \end{cases} \tag{2}$$

where the action A^i is an external instruction that guides focus adjustments. The LMM model L refines the textual focus based on the current focus f^i and action A^i, while the visual expert E updates the image region based on the refined text cue T^{i+1}. This iterative process allows A to dynamically guide both visual and textual update, enabling dynamic focus adjustment.

To simulate human focus-shifting actions, we designed an action space that guides the LMM and visual expert in coordinated interaction. Unlike methods that rely on LLMs to generate textual rationales [27], our actions are crafted to emulate human visual behavior rather than purely semantic reasoning.

◇ ***A1****: Semantic Focus*. This action identifies contextually relevant visual targets based on semantic information (e.g., query-related objects), simulating conjunction search and endogenous orienting [2] to match targets according to semantic cues.

◇ ***A2****: Semantic Scatter*. The action enlarges the current focused area, avoiding the loss of key information caused by overly precise focusing, simulating the divergence action of the human eye's focus.

3.2.2 Focus Tree Search

We propose Focus Tree Search (FTS), a Monte Carlo Tree Search (MCTS) [4, 14] based visual search method that navigates through visual spaces with *focus* nodes. FTS efficiently combines focus adjustment actions, balancing exploration and exploitation to construct a focus tree that retains key visual information while eliminating irrelevant details. In our approach, FTS incrementally builds a focus tree, where each node represents a *focus* state f, and each edge corresponds to a focus-swift action that transitions to a new focus region. To efficiently guide the Large Multimodal Model (LMM) and visual expert in exploring the most promising nodes, we maintain a state-action value function $Q : f \times \mathcal{A} \mapsto \mathbb{R}$, which estimates the *expected future reward* of performing action a in *focus* state f .

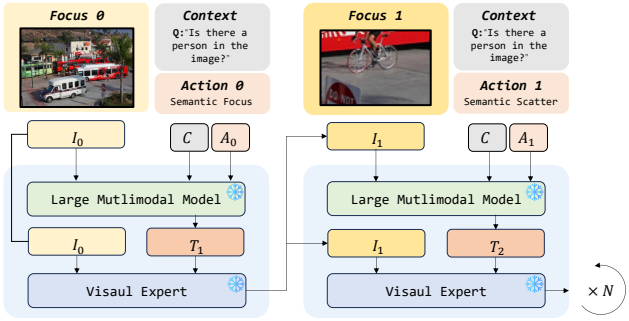


Figure 3. An illustration of Focus Adjuster of DyFo.

Selection Phase. The algorithm starts at the root node (initial focus state f^0) and iteratively selects the next node at each tree level by balancing exploration and exploitation. This phase continues until a leaf node is reached. To achieve the balance between exploration (visiting less-visited nodes) and exploitation (visiting high-value nodes), we use the well-established *Upper Confidence bounds applied to Trees (UCT)* [14] to select each child node. Specifically, at node f, we select an action by considering both the Q-value (for exploitation) and uncertainty (for exploration):

$$a^* = \arg\max_{a \in A(f)} \left[Q(f, a) + w \sqrt{\frac{\ln N(f)}{N(c(f, a))}} \right], \tag{3}$$

where $N(f)$ represents the number of visits to node f in previous iterations, and $c(f, a)$ is the child node resulting from applying action a in state f. The second term increases with uncertainty (fewer visits to the child node). The weight w controls the exploration-exploitation trade-off.

Expansion Phase. During this phase, new child nodes are added to the selected leaf node by randomly sampling unexplored actions. If the leaf node is a terminal node (i.e., the search count has reached its maximum), the expansion phase is skipped, and start backpropagation immediately.

Backpropagation Phase. Once a terminal state is reached in the above phase, the search path from the root node to the terminal node is obtained. At this point, rewards are backpropagated along the path to update the Q value for each state-action pair on the path. Specifically, $Q(f, a)$ is up-

Dataset	Setting	Model	Method	Accuracy↑	Precision	Recall	F1 Score↑
MSCOCO	*Random*	LLaVA1.5	Regular	90.07	94.19	85.40	89.58
			DyFo	**91.40**	90.86	92.07	**91.46**
		Qwen2-VL	Regular	90.07	97.55	82.20	89.22
			DyFo	**92.13**	96.95	87.00	**91.71**
	Popular	LLaVA1.5	Regular	87.27	88.71	85.40	87.02
			DyFo	**88.47**	85.88	92.07	**88.87**
		Qwen2-VL	Regular	88.57	94.06	82.33	87.81
			DyFo	**89.97**	92.13	87.40	**89.70**
	Adversarial	LLaVA1.5	Regular	81.83	79.86	85.13	82.41
			DyFo	**83.40**	78.40	92.20	**84.74**
		Qwen2-VL	Regular	85.67	88.54	81.93	85.11
			DyFo	**86.77**	86.55	87.07	**86.81**
A-OKVQA	*Random*	LLaVA1.5	Regular	86.80	82.17	94.00	87.69
			DyFo	**88.17**	82.51	96.87	**89.11**
		Qwen2-VL	Regular	89.33	94.49	83.53	88.68
			DyFo	**91.40**	94.11	88.33	**91.13**
	Popular	LLaVA1.5	Regular	82.87	76.88	94.00	84.58
			DyFo	**83.70**	76.70	96.80	**85.59**
		Qwen2-VL	Regular	88.20	92.01	83.67	87.64
			DyFo	**89.27**	90.18	88.13	**89.14**
	Adversarial	LLaVA1.5	Regular	72.57	65.80	94.00	77.41
			DyFo	**73.47**	65.99	96.87	**78.50**
		Qwen2-VL	Regular	81.07	79.61	83.53	81.52
			DyFo	**82.20**	78.65	88.40	**83.24**
GQA	*Random*	LLaVA1.5	Regular	87.67	82.66	95.33	88.54
			DyFo	**88.50**	82.91	97.00	**89.40**
		Qwen2-VL	Regular	78.50	98.09	58.13	73.00
			DyFo	**86.60**	95.75	76.60	**85.11**
	Popular	LLaVA1.5	Regular	79.90	72.82	95.40	82.60
			DyFo	**80.27**	72.70	96.93	**83.09**
		Qwen2-VL	Regular	78.07	96.67	58.13	72.61
			DyFo	**84.77**	91.36	76.80	**83.45**
	Adversarial	LLaVA1.5	Regular	73.50	66.36	95.33	78.25
			DyFo	**74.07**	66.50	97.00	**78.90**
		Qwen2-VL	Regular	75.90	90.09	58.20	70.72
			DyFo	**81.97**	85.39	77.13	**81.05**

Table 1. Results on POPE. *Regular* denotes LMM directly answer question based on input image. DyFo refers to our method. The best performances within each setting are **bolded**.

dated by aggregating the rewards from all subsequent steps of node f, as follows:

$$Q(f, a) = \Sigma_{f^i \in N_f} R_{f^i}, \tag{4}$$

where N_f is the set of child nodes of node f, and R_{f^i} is the reward value corresponding to each child node.

To encourage the LMM and visual expert to filter out irrelevant content while avoiding the introduction of biases (e.g., the visual expert locating the wrong target or the LMM generating visual hallucinations), we use consistency between the LMM and the visual expert as a criterion, with the effective area of the node serving as the search reward:

$$R_{f^i} = \mathbb{I}_{\{I=T\}} \cdot \frac{s_{f^i}}{s_o}, \tag{5}$$

where R_{f^i} represents the value of the reward function. $\mathbb{I}_{\{I=T\}}$ is an indicator function that takes a value of 1 if the image I associated with node f^i is semantically consistent with the text T, and 0 otherwise. $\frac{s_{f^i}}{s_o}$ denotes the effective area ratio of node f^i, where s_{f^i} is the effective area and s_o is the area of the original input image.

Multi-grained Voting Once the above phases are complete and the focus tree is constructed, our goal is to fully utilize the essential visual information within this tree to achieve more precise fine-grained visual understanding. Inspired by the self-consistency method [47], we adopt a multi-node voting approach to obtain the final answer. Specifically, the final result is obtained through a weighted voting scheme across different nodes, where each node f provides a prediction $L(f)$, and the weight for each node is the reward R_f it has received. Formally, the final answer $\mathcal{A}$ is derived as:

$$\mathcal{A} = \arg\max_{L(f)} \sum_{f \in \mathcal{T}_f} R_f \cdot L(f) \tag{6}$$

where $\mathcal{T}_f$ is the focus tree composed. This voting mechanism ensures that the model avoids an overemphasis on de-

Sampling	Metrics	DyFo-Q	DyFo-L	Osprey-7B	Ferret-7B	Shikra-7B	LLaVA-1.5	Qwen2-VL	InstructBLIP	MiniGPT4	mPLUG-Owl
Random	Accuracy ↑	**92.13**	91.40	89.47	90.24	86.90	88.73	90.07	88.57	79.67	53.97
	Precision	96.95	90.91	93.40	97.72	94.40	88.89	97.55	84.09	78.24	52.07
	Recall	87.00	92.00	84.93	83.00	79.26	88.53	82.20	95.13	82.20	99.60
	F1 Score ↑	**91.71**	91.45	88.97	89.76	86.19	88.71	89.22	89.27	80.17	68.39
	Yes (%)	48.65	50.60	45.47	43.78	43.26	49.80	43.92	56.57	52.53	95.63
Popular	Accuracy ↑	**89.97**	88.57	87.83	84.90	83.97	85.83	88.57	82.77	69.73	50.90
	Precision	92.13	86.09	89.94	88.24	87.55	83.91	94.06	76.27	65.86	50.46
	Recall	87.40	92.00	85.20	80.53	79.20	88.67	82.33	95.13	81.93	99.40
	F1 Score ↑	**89.70**	88.95	87.50	84.21	83.16	86.22	87.81	84.66	73.02	66.94
	Yes (%)	48.59	53.43	47.37	45.63	45.23	52.83	46.72	62.37	62.20	98.57
Adversarial	Accuracy ↑	**86.77**	83.43	85.33	82.36	83.10	72.10	85.67	65.17	79.20	50.67
	Precision	86.55	78.41	85.43	83.60	85.60	74.69	88.54	65.13	61.19	50.34
	Recall	87.07	92.27	85.20	80.53	79.60	88.34	81.93	95.13	82.93	99.33
	F1 Score ↑	**86.81**	84.78	85.31	82.00	82.49	80.94	85.11	77.32	70.42	66.82
	Yes (%)	49.65	58.83	49.87	48.18	46.50	59.14	45.82	73.03	67.77	98.67

Table 2. Results on POPE-COCO benchmark comparing our method with existing LMMs. DyFo-Q is our method with Qwen2-VL, DyFo-L is our method with LLaVA-v1.5. The best performances within each setting are **bolded**.

	Attribute (%)	Spatial (%)	Overall (%)
Human	98.26	100.00	98.95
Random Guess	26.73	50.00	35.99
Open-source end-to-end LMMs			
BLIP2 [19]	26.95	53.94	37.69
MiniGPT-4 [64]	30.43	50.00	38.22
LLaVA [24]	23.47	53.94	35.59
InstructBLIP [5]	25.21	47.36	34.02
Otter [16]	26.95	56.57	38.74
LLaVA-1.5 [23]	43.47	56.57	48.68
Qwen2-VL [45]	74.78	68.42	72.25
Commercial chatbot systems			
Bard [8]	31.30	46.05	37.17
Gemini Pro [6]	40.86	59.21	48.16
GPT-4V [28]	51.30	60.52	54.97
LLM tool-using pipelines			
MM-React [54]	34.78	51.31	41.36
VisualChatGPT [50]	30.43	48.68	37.69
Visprog [10]	31.30	56.57	41.36
Visual search			
SEAL [51]	74.78	76.31	75.39
DyFo-L (ours)	62.61	53.95	59.16 (+10.48)
DyFo-Q (ours)	**80.00**	**82.89**	**81.15** (+8.90)

Table 3. Results on V* Bench comparing our method with existing LMMs.

tails that could lead to the loss of critical global hint while still preserving the importance of relevant details.

4. Experiments

4.1. Experimental Settings

Datasets & Evaluation Metrics

POPE (Polling-based Object Probing Evaluation) [20] offers a general benchmark for assessing visual hallucination in LMMs. This evaluation framework systematically tests model accuracy in identifying object presence within images by presenting balanced queries of existent and non-existent objects (50% each). The benchmark includes three distinct sampling strategies for constructing negative samples: *random*, *popular*, and *adversarial*. Specifically, in the *random* setting, absent objects are chosen randomly; in the *popular* setting, they are selected from a pool of frequently appearing objects; and in the *adversarial* setting, objects are chosen based on common co-occurrences yet are absent from the image. Moreover, the POPE benchmark integrates data from three sources—MSCOCO [22], A-OKVQA [37], and GQA [12]. Across these sampling strategies, 500 images are drawn from each dataset, with 6 queries generated per image, resulting in a total of 27,000 query-answer pairs. Finally, evaluation is conducted using four core metrics: Accuracy, Precision, Recall, and F1 score.

V* Bench [51] is a challenging benchmark designed to assess large vision-language models (LMMs) on fine-grained detail recognition in high-resolution images. Based on the SA-1B dataset, it includes 191 images with an average resolution of 2246×1582 and covers two key tasks: attribute recognition (115 samples) and spatial relationship reasoning (76 samples). Attribute recognition evaluates models on identifying detailed object attributes like color, material, and shape, while spatial relationship reasoning testes models on determining relative spatial positions among objects. Accuracy is used as the evaluation metric.

Baseline Methods We assessed the effectiveness of our proposed DyFo on two representative LMMs: LLaVA-1.5, a fixed-resolution model, and Qwen2-VL, a variable-resolution model, using their 7B versions unless otherwise specified. To further validate the competitiveness of our approach, we compared it with recent approaches aimed at enhancing the fine-grained visual understanding capabilities of LMMs [5, 16, 19, 55, 64]. These include models that improve regional comprehension with additional data and training phrase [3, 57, 59], LLMs employing multi-tool combinations to extend visual capabilities [10, 50, 54],

Model	A_1	A_2	Attribute(%)↑	Spatial(%)↑
LLaVA-v1.5	✓		57.50	55.50
		✓	45.46	52.61
	✓	✓	62.61	53.95
Qwen2-VL	✓		76.78	70.25
		✓	75.67	69.50
	✓	✓	80.00	82.89

Table 4. Exploration of action space combination on V* Bench.

Model	Method	Accuracy↑	F1 Score↑
Random	Visual Expert	88.87	88.53
	LLaVA-v1.5	88.73	88.71
	Qwen2-VL	90.07	89.22
	DyFo-L	91.40	91.45
	DyFo-Q	**92.13**	**91.71**
Popular	Visual Expert	85.47	85.58
	LLaVA-v1.5	85.83	86.22
	Qwen2-VL	88.57	87.81
	DyFo-L	88.57	88.95
	DyFo-Q	**89.97**	**89.70**

Table 5. Exploration of the impact of visual expert on POPE-COCO with random and popular setting.

closed-source commercial chatbot systems (GPT-4V [28], Gemini Pro [41], Bard[8]), and visual search methods tailored for LMMs equipped with localization and heatmap output modules [51].

Implementation Details In our experiments, to minimize the operational costs associated with visual experts, we utilize Lang-Segment-Anything [26] as the sole visual expert. This system efficiently provides detection boxes and masks based on input text through batch processing, which responds faster and easy to deployment. All experiments were conducted on two NVIDIA L40 GPUs.

Category	Method	Search Length ↓
Space Design	Uniform Dividing Policy	4.90
	Dynamical Focus	**3.20**
Search Algorithm	A* Search	4.33
	MCTS	**3.20**
	BFS	5.65
	DFS	6.15

Table 6. Comparison of search length across different search algorithms and search space design on V* Bench. The best result is in **bold**.

4.2. Experimental Results

POPE Benchmark Results We first evaluate the potential of our method to enhance the general visual capabilities of LMMs on the POPE benchmark across three datasets (COCO, AOKVQA, GQA) and three sampling strategies (random, popular, adversarial). The results are summarized in Table 1. A key observation is that our proposed `Dynamic Focus` (`DyFo`) consistently improves the performance of both fixed-resolution and dynamic-resolution models across all datasets and settings. These results suggest that `DyFo` effectively reduces visual hallucinations by emphasizing relevant visual content and minimizing the impact of irrelevant factors, thereby enhancing the visual understanding of both fixed-resolution and variable-resolution LMMs. To further assess the competitiveness of our approach, we compare it with several existing methods aimed at improving fine-grained visual understanding in LMMs. Based on the results in Table 2, `DyFo` consistently outperforms the comparison methods across all three sampling settings (random, popular, and adversarial) in the COCO dataset. These results demonstrate that `DyFo`, by combining focus tree search and expert collaboration, is able to effectively filter out irrelevant or misleading visual content, reducing hallucinations.

Results on V* Bench The V* Bench evaluation set focuses on fine-grained visual understanding tasks with high-resolution images and small object targets, posing challenges for LMMs in focusing details. As shown in Table 3, most large multimodal models (LMMs) perform near random levels, and current tool-based methods also struggle to achieve meaningful results. Notably, as a visual search method requiring no additional adaptation or training of LMMs, `DyFo` surpasses the visual search model SEAL, which fine-tunes LMMs specifically for the visual search task in both architecture and paradigm. This demonstrates the potential of flexible visual search methods that integrate visual experts with LMMs.

4.3. Ablation Study

Effect of Action Space on Visual Localization We further analyzed the effect of action space design on LMM performance in fine-grained visual understanding. Table 4 shows that using only single actions results in performance declines, while combining even two actions brings measurable gains.

Impact of visual expert To evaluate the advantages of our method compared to standalone visual experts and independent LMMs, we compared our method with only visual expert on POPE-COCO with three settings. Results in Table 5 indicate that our method consistently surpasses standalone visual expert. This suggests that combining LMMs with visual experts not only benefits from the experts' visual

capabilities but also enables LMMs to mitigate biases and limitations in visual experts, resulting in mutual enhancement—consistent with our initial analysis.

Efficiency Study To further evaluate the efficiency of our search space and algorithm design, we compared them with respective counterparts using the V* Bench dataset. The search objective is to locate the node with the highest reward value within fixed max depth of 5. The search process stops either upon reaching the maximum number of iterations or identifying the target node, as shown in Table 6. DyFo 's search space design requires fewer search steps than the Uniform Dividing Policy taken by SEAL [51]. The result shows that our method may avoid ambiguous search steps that uniform division method have. To assess whether Monte Carlo Tree Search (MCTS) could search efficiently by balancing exploration and exploitation, we compared it with A*, BFS, and DFS search algorithms. Experimental results indicate that MCTS achieves the highest search efficiency, which we attribute to its unique capability for reward function modeling. This feature allows it to avoid greedy traps by gradually approximating the true reward distribution. Notably, while A* theoretically guarantees the shortest path under admissible conditions, in practice, we found that finding appropriate heuristics to meet these conditions is challenging, especially when the target node is uncertain. We believe this limitation causes A* to perform similarly to greedy algorithms in our setting.

Case Study Figures 4 and 5 present case studies from the POPE and V* Bench datasets. In the low-resolution example in Figure 4 , objects like the "*wall* " and "*baseball bat* " are off-center, partially obscured, and occupy only a small portion of the area. This makes it challenging for the LMM to accurately focus on these elements amidst a large amount of visual content, leading to hallucination errors. DyFo dynamically adjusts the focal area (indicated by the red box), allowing the LMM to filter out irrelevant content and mitigate inherent hallucination issues. In the high-resolution example shown in Figure 5, the challenge of focusing on key objects is even greater: objects like the "*glove*" and "*dove*" occupy less than $1/50$ of the image area, meaning that nearly 98% of the visual input is irrelevant or distracting to the LMM. Through dynamic focusing, our method minimizes interference from irrelevant content and retains only the most critical visual information, improving performance on fine-grained visual understanding tasks.

5. Conclusion

We propose DyFo (Dynamic Focus), a training-free dynamic focus visual search method that enhances fine-grained visual understanding in large multimodal models (LMMs). By simulating human-like focusing adjustments, our method enables LMMs to selectively amplify critical visual information while filtering out irrelevant input, improving the model's ability to handle complex visual tasks with higher accuracy. Experimental results demonstrate that DyFo reduces visual hallucinations and enhances performance on fine-grained visual understanding tasks across various datasets suggesting that DyFo provides a promising approach to improving the visual capabilities of LMMs.

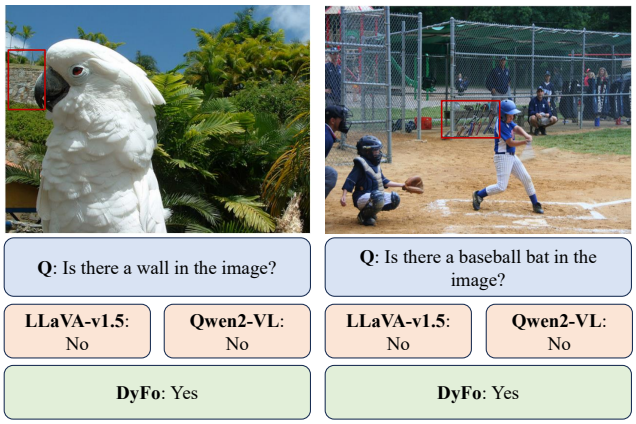


Figure 4. Comparison between the responses of LLaVA-v1.5, Qwen2-VL and our method DyFo on POPE cases. The final focus region is highlighted in the image using red bounding boxes.



Figure 5. Comparison between the responses of LLaVA-v1.5, Qwen2-VL and our method DyFo on V* Bench several cases. The final focus region is highlighted in the image using red bounding boxes.

Acknowledgements

This work was supported by the grants from the National Natural Science Foundation of China (61925201, 62132001, 62432001, 62373043), Beijing Natural Science Foundation (L247006, 4252020) and sponsored by Tencent Basic Platform Technology Rhino-Bird Focused Research Program.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1

[2] Andrea Berger, Avishai Henik, and Robert Rafal. Competition between endogenous and exogenous orienting of visual attention. *Journal of experimental psychology: General*, 134 (2):207, 2005. 4

[3] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1, 2, 6

[4] Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *Computers and Games: 5th International Conference, CG 2006, Turin, Italy, May 29-31, 2006. Revised Papers 5*, pages 72–83. Springer, 2007. 4

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Proceedings of the Advances in Neural Information Processing Systems*, 36, 2024. 1, 6

[6] Google DeepMind. Gemini, 2023. 6

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[8] Google. Bard, 2023. 6, 7

[9] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 1

[10] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 6

[11] Sun Haoran, Wang Yang, Liu Haipeng, and Qian Biao. Fine-grained cross-modal fusion based refinement for text-to-image synthesis. *Chinese Journal of Electronics*, 32(6): 1329–1340, 2023. 1

[12] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. 6

[13] EB Huey. The psychology and pedagogy of reading (reprinted 1968), 1908. 1

[14] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *Machine Learning: ECML 2006: 17th European Conference on Machine Learning Berlin, Germany, September 18-22, 2006 Proceedings 17*, pages 282–293. Springer, 2006. 4

[15] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 3

[16] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 6

[17] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024. 1

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, pages 12888–12900, 2022. 2

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2, 6

[20] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 6

[21] Yonglin Li, Jing Zhang, Xiao Teng, and Long Lan. Refsam: Efficiently adapting segmenting anything model for referring video object segmentation. *arXiv preprint arXiv:2307.00997*, 2023. 3

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 6

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Proceedings of the Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 6

[25] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 3

[26] Luca Medeiros. Language segment-anything, 2024. 2, 7

[27] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. *arXiv preprint arXiv:2311.17076*, 2023. 4

[28] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6, 7

[29] Marius V Peelen and Sabine Kastner. A neural basis for real-world visual search in human occipitotemporal cortex. *Proceedings of the National Academy of Sciences*, 2011. 1

[30] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 1

[31] Roy M Pritchard, Woodburn Heron, and DO Hebb. Visual perception approached by the method of stabilized images. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 14(2):67, 1960. 1

[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 *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021. 3

[33] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 2

[34] Keith Rayner. Eye movements and cognitive processes in reading, visual search, and scene perception. In *Eye Movement Research*, pages 3–22. North-Holland, 1995. 1

[35] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 3

[36] Michele Rucci and Martina Poletti. Control and functions of fixational eye movements. *Annual review of vision science*, 1(1):499–518, 2015. 1

[37] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 6

[38] Melanie Sclar, Gaston Bujia, Sebastian Vita, Guillermo Solovey, and Juan Esteban Kamienkowski. Modeling human visual search: A combined bayesian searcher and saliency map approach for eye movement guidance in natural scenes. In *NeurIPS Workshop SVRHM*, 2020. 2

[39] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[40] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016. 2

[41] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 7

[42] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 2006. 1, 2

[43] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. Evaluation and analysis of hallucination in large vision-language models, 2023. 1

[44] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation, 2024. 1

[45] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2, 6

[46] Sisi Wang, Stanislas Huynh Cong, and Geoffrey F Woodman. Statistical learning speeds visual search: More efficient selection, or faster response? *Journal of Experimental Psychology: General*, 2023. 1

[47] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. 5

[48] Jeremy M Wolfe. Visual search: How do we find what we are looking for? *Annual review of vision science*, 2020. 1

[49] Jeremy M Wolfe, Melissa L-H Võ, Karla K Evans, and Michelle R Greene. Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, 2011. 1

[50] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual ChatGPT: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 2, 6

[51] 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. 2, 3, 6, 7, 8

[52] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023. 3

[53] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 193–202, 2020. 2

[54] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu,

Michael Zeng, and Lijuan Wang. mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 2, 6

[55] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 6

[56] Zhaoda Ye, Xiangteng He, and Yuxin Peng. Unsupervised cross-media hashing learning via knowledge graph. *Chinese Journal of Electronics*, 31(6):1081–1091, 2022. 1

[57] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 1, 2, 6

[58] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024. 1

[59] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28202–28211, 2024. 6

[60] Mengmi Zhang, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman. Finding any waldo with zero-shot invariant and efficient visual search. *Nature communications*, 2018. 2

[61] Rui Zhang, Cong Xie, and Liwei Deng. A fine-grained object detection model for aerial images based on yolov5 deep neural network. *Chinese Journal of Electronics*, 32(1):51–63, 2023. 1

[62] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model, 2024. 2

[63] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything, 2023. 3

[64] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 6

[65] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*, 2024. 2