SIFThinker: Spatially-Aware Image Focus for Visual Reasoning

Zhangquan Chen[1*], **Ruihui Zhao**[3], **Chuwei Luo**[3], **Mingze Sun**[1], **Xinlei Yu**[4], **Yangyang Kang**[2,3†], **Ruqi Huang**[1†]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University, China
[2]Zhejiang University, China
[3]ByteDance, China
[4]National University of Singapore, Singapore
yangyangkang@bytedance.com, ruqihuang@sz.tsinghua.edu.cn

Abstract

Current multimodal large language models (MLLMs) still face significant challenges in complex visual tasks (e.g., spatial understanding, fine-grained perception). Prior methods have tried to incorporate visual reasoning, however, they fail to leverage attention correction with spatial cues to iteratively refine their focus on prompt-relevant regions. In this paper, we introduce *SIFThinker*, a spatially-aware "think-with-images" framework that mimics human visual perception. Specifically, *SIFThinker* enables attention correcting and image region focusing by interleaving depth-enhanced bounding boxes and natural language. Our contributions are twofold: First, we introduce a reverse-expansion-forward-inference strategy that facilitates the generation of interleaved image-text chains of thought for process-level supervision, which in turn leads to the construction of the **SIF-50K** dataset. Besides, we propose **GRPO-SIF**, a reinforced training paradigm that integrates depth-informed visual grounding into a unified reasoning pipeline, teaching the model to dynamically correct and focus on prompt-relevant regions. Extensive experiments demonstrate that *SIFThinker* outperforms state-of-the-art methods in spatial understanding and fine-grained visual perception, while maintaining strong general capabilities, highlighting the effectiveness of our method.

Code — https://github.com/zhangquanchen/SIFThinker
Extended version — https://arxiv.org/pdf/2508.06259

Introduction

Visual understanding is a fundamental task in computer vision, enabling machines to perceive, interpret, and interact with their surroundings (Wolfe and Horowitz 2017; Guo et al. 2016; Palmeri and Gauthier 2004). Traditional methods typically process the entire image in a uniform manner (Chen et al. 2024b, 2025b; Jiao et al. 2025; Shao et al. 2024a; Zhang et al. 2024), without considering the dynamic attention shifts and the underlying spatial awareness. However, an RGB image is inherently the 2D projection of 3D world, and human perception is dynamic attention shifts rooted in 3D awareness. For example, when prompted to identify the color of shirt worn by the person behind the tree, humans do not perceive the scene in a single step. They first locate a coarse region of interest within their field of view. Then, they progressively focus on the tree while considering its 3D spatial relationship to the surroundings in mind. Finally, they correct their attention to the area behind the tree to determine the color of the shirt worn by the person.

The above observations indeed reveal two critical points in designing a human-like visual understanding framework – 1) **Dynamic visual perception**, which enables attention correction and facilitates focusing on prompt-relevant regions throughout the reasoning process; and 2) **3D spatial awareness**, which grounds visual perception within a spatial context. Moreover, *these two points can be naturally integrated into a unified framework*.

For the former, dynamic visual perception was initially pursued through reasoning frameworks. Early efforts, such as V* (Wu and Xie 2024), VisCoT (Shao et al. 2024b), and VisRL (Chen, Luo, and Li 2025), incorporated visual inputs into the reasoning paradigm in a stepwise manner: first predicting bounding boxes, then performing further inference based on the cropped image. This kind of fragmented method often severs the continuity of the reasoning chain, leading to weaker interpretability and incoherent anchoring of visual regions. More recent think-with-images methods, such as Cogcom (Qi et al. 2024) and ChatGPT-o3 (OpenAI 2025), utilize specialized-model-based or tool-based adaptive zooming to simulate dynamic attention. However, these methods are not intrinsic and heavily rely on external capabilities. In contrast, *SIFThinker intrinsically supports coherent image-text interleaved reasoning by simulating human-like visual attention correction and focusing*.

For the latter, the limitation primarily arises from the insufficient modeling of spatial information in MLLMs, i.e., most of them are pre-trained on RGB images paired with textual data without explicit spatial cues. Humans, on the

[*]Work was done during the internship at ByteDance.
[†]Corresponding authors.

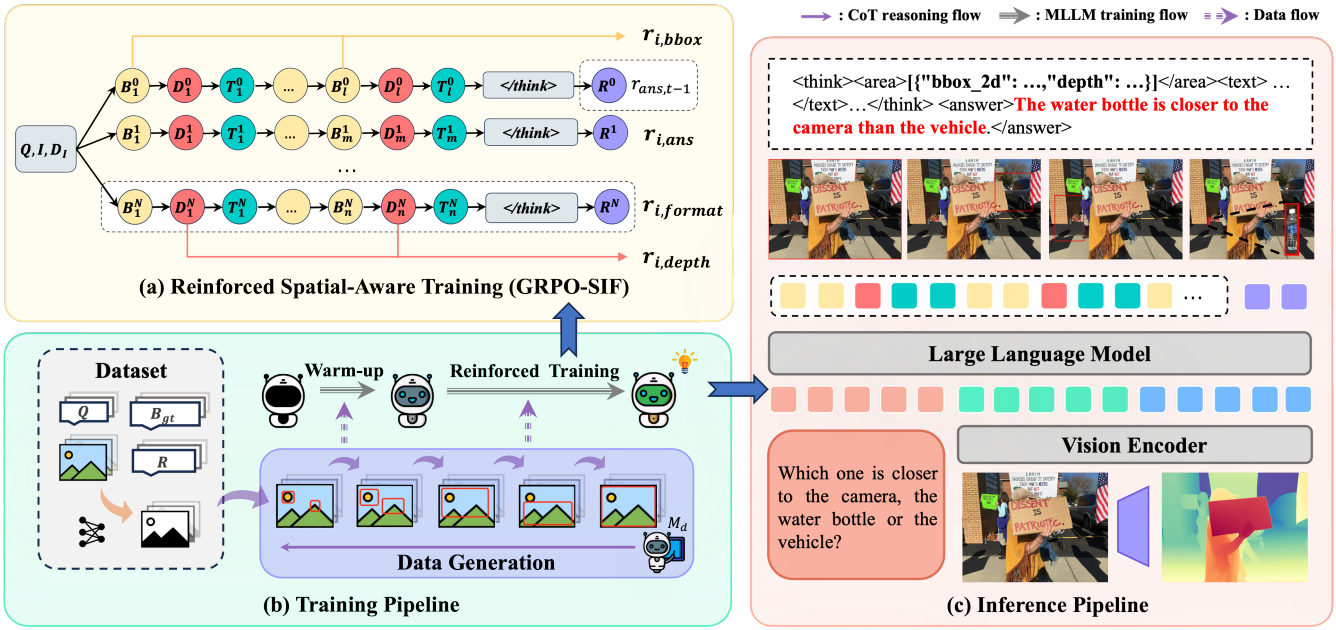

Figure 1: The schematic illustration of *SIFThinker*. (a) We propose a spatially-aware image focus paradigm, in which four novel reward functions are introduced under the RL framework. (b) The training pipeline of *SIFThinker* is illustrated. It begins with a warm-up stage, followed by GRPO-SIF as described in (a). (c) illustrates the inference pipeline of our method.

other hand, are often capable of inferring 3D relationships from a single 2D image, enabling a deeper understanding of visual scenes. Prior works such as SpatialBot (Cai et al. 2024) and SSR (Liu et al. 2025b) attempt to equip models with 3D perception via data-driven or tool-based approaches. However, In real-world interactive scenarios, attention is seldom global. In contrast, *we typically perceive the spatial information (e.g., depth) of the prompt-relevant regions rather than the entire image*.

Moreover, previous works often fail to *integrate fine-grained visual perception with spatial awareness*, and typically separate visual grounding from answer generation (e.g., VLM-R1 (Shen et al. 2025)). In contrast, we argue that 3D perception serves as a fundamental basis for deeper visual understanding. Besides, grounding-conditioned reasoning facilitates more accurate answer derivation, while result-level supervision can guide the correction of bounding boxes. Thus, there is a need for a spatially-aware, visually grounded reasoning process that offers a more coherent, unified, and effective framework for visual understanding.

To this end, we propose an adaptive image focus mechanism with 3D awareness, enabling the model to selectively and dynamically perceive the focused region with depth throughout the reasoning process. That is, we *integrate visual perception and spatial awareness into a unified grounded reasoning pipeline*.

Our method consists of three components: data generation, warm-up, and reinforcement learning (RL). In the data generation stage, we propose a novel image–text interleaved chain-of-thought (CoT) generation scheme that simulates spatially-aware visual focus. The constructed data support the warm-up stage via supervised fine-tuning (SFT), guiding the model to adopt structured interleaved reasoning. In the reinforcement learning stage, we go beyond outcome-only rewards and integrate format, depth estimation, region grounding, and answer prediction for each rollout.

Our main contributions can be summarized as follows.

- We present *SIFThinker*, the first framework to incorporate an adaptive focus mechanism with 3D awareness, enabling spatially grounded visual reasoning.
- We introduce a novel reverse-expansion–forward-inference strategy for constructing interleaved CoTs, and release **SIF-50K** dataset for process-level supervision.
- We propose **GPRO-SIF**, which applies reinforcement learning for spatially-aware image focus training: 1) for visual grounding, we propose $HIoU$, a hierarchical design that better accommodates multiple objects, along with an effective reward formulation for bounding box correction; 2) for spatial awareness, we embed depth estimation into the model's autoregressive structure; 3) we further design a progressive reward to encourage performance improvement across training epochs.
- Extensive experiments demonstrate that *SIFThinker* outperforms prior state-of-the-art (SOTA) methods in both fine-grained visual perception and spatial intelligence, while maintaining stable general capabilities.

Related Work

Visual Chain-of-Thought Reasoning

Recent studies have shown that step-by-step reasoning through in-context learning can significantly enhance

the performance of large language models (LLMs). Accordingly, several approaches have emerged that aim to strengthen the visual reasoning capabilities of multi-modal large language models by introducing chain-of-thought strategies. These approaches can be categorized into three types: (T1) Pure-Text-Thought Methods: (Chen et al. 2025a; Thawakar et al. 2025; Ji et al. 2024; Hu et al. 2024; Shen et al. 2025; Bai et al. 2025b) elicit purely textual CoT reasoning within MLLMs for visual reasoning tasks inspired by (Guo et al. 2025b). They employ reinforcement learning to guide the generation process towards final answers without explicitly incorporating intermediate visual signals. (T2) Intermediate-Thought Methods: (Liu et al. 2025a; Shao et al. 2024b; Chen, Luo, and Li 2025; Wang et al. 2024) first generate fine-grained visual cues (e.g, bounding boxes, spatial coordinates, or segmentation masks), CoT reasoning is then conducted based on these fine-grained visual cues. (T3) Multi-Modal-Thought Methods: Some recent methods aim to integrate visual-textual reasoning more tightly into the model's thought process. Proprietary systems such as ChatGPT-o3 (OpenAI 2025) demonstrate the ability to "think-with-images" by dynamically invoking external image tools. Similarly, (Li et al. 2025) enables visual thinking by generating visual traces of reasoning. (Su et al. 2025; Wu et al. 2025; Zheng et al. 2025) optimize tool-usage ability via reinforcement learning. Moreover, (Zhang et al. 2025) iteratively crops the image based on generated bounding boxes to extract new visual cues for reasoning. (Fan et al. 2025) takes a more direct approach by generating reasoning chains that interleave natural language and explicit bounding boxes.

However, existing methods still exhibit several limitations: (T1) they rely heavily on textual reasoning, neglecting the dynamic visual attention shifts during the reasoning process; (T2) they produce less interpretable and often incoherent reasoning chains; and (T3) Some depend on external tools, specialized detection models or unstable image generation, while others overlook intermediate visual signals, relying solely on outcome-based supervision. Therefore, there is a need for an adaptive and coherent intrinsic method that enables visual-grounded reasoning—allowing MLLMs not only think "about" images, but also *dynamically focus, and correct their visual attention across image regions in a human-like manner.*

Spatial Intelligence

Existing MLLMs (Wu et al. 2024; Driess et al. 2023; Li et al. 2023b; Chen et al. 2022) are primarily trained on RGB images paired with textual data, which inherently lacks 3D spatial information. As a result, they demonstrate limited performance on tasks requiring spatial reasoning. To address this limitation, recent efforts such as SpatialRGPT (Cheng et al. 2024) and SpatialVLM (Chen et al. 2024a) have enhanced the spatial reasoning capabilities of MLLMs by constructing specialized spatially-oriented question-answering datasets and fine-tuning models accordingly. To further emphasize integrated reasoning capabilities, SSR (Liu et al. 2025b) incorporates depth images as additional inputs, while SpatialBot (Cai et al. 2024) leverages depth estimation tools to acquire spatial priors in key perceptual regions. However, these spatial perception approaches focus solely on reasoning, without achieving a deep integration with visual grounding—*two processes that are fundamentally interdependent in human visual perception.*

Algorithm 1: CoT Completion

Input: One sample from the source dataset including question Q, image I, normalized b-boxes B_{gt}, and response R. Doubao-1.5-vision-pro model $\mathcal{M}_d$. DepthAnythingV2 model $\mathcal{M}_{DA}$.
Output: Completed reasoning chain R_{cot}.

$B_0 \leftarrow B_{gt}; \mathcal{B} \leftarrow \{B_0\}; ET \leftarrow False; \mathcal{T} = 0.2$
for $t \leftarrow 1$ **to** $K = 5$ **do**
  $S \leftarrow K - t + 1; B_t \leftarrow \emptyset$
  foreach *b-box* $b = (x_1, y_1, x_2, y_2) \in B_{t-1}$ **do**
    $\delta_{x_1} \leftarrow \frac{x_1}{S}, \delta_{x_2} \leftarrow \frac{1-x_2}{S}$
    $\delta_{y_1} \leftarrow \frac{y_1}{S}, \delta_{y_2} \leftarrow \frac{1-y_2}{S}$
    $b' \leftarrow (x_1 - \delta_{x_1}, y_1 - \delta_{y_1}, x_2 + \delta_{x_2}, y_2 + \delta_{y_2})$
    $B_t \leftarrow B_t \cup \{b'\}$
  for $(b_i, b_j) \in B_t$ ***where*** $IOU(b_i, b_j) > 0$ **do**
    $b_{\text{merged}} \leftarrow \left(\min_{k\in\{i,j\}} x_1^k,\ \min_{k\in\{i,j\}} y_1^k,\ \max_{k\in\{i,j\}} x_2^k,\ \max_{k\in\{i,j\}} y_2^k\right)$
    $B_t \leftarrow \Big(B_t \setminus \{b_i, b_j\}\Big) \cup \{b_{\text{merged}}\}$
  $\mathcal{B} \leftarrow \mathcal{B} \cup \{B_t\}$
  if $|B_t| = 1$ ***and not*** ET **then**
    $K \leftarrow t + 2, ET \leftarrow True$
for $i \leftarrow 1$ **to** *random choice from* $\{0, 1, 2\}$ **do**
  // $\mathcal{U}_I(\mathcal{B})$:`B-boxes outside` $\mathcal{B}$; $\mathcal{R}_{\mathcal{T}}(\cdot)$: `Random selection with an area difference no more than` $\mathcal{T}$
  $B_{K+i} = \mathcal{R}_{\mathcal{T}}(\mathcal{U}_I(\bigcup_{j=K}^{K+i-1} B_j))$
  $\mathcal{B} \leftarrow \mathcal{B} \cup \{B_{K+i}\}$
$\mathcal{B}_{rev} \leftarrow \text{reverse}(\mathcal{B}); I_{seq} \leftarrow \text{DrawBBoxes}(I, \mathcal{B}_{rev})$
$D_I \leftarrow \mathcal{M}_{DA}(I)$
$R_{cot} \leftarrow \mathcal{M}_d\big(Q,\ R,\ D_I,\ (I_{seq},\ \mathcal{B}_{rev})\big)$
return R_{cot}

Methodology

As shown in Fig. 1, *SIFThinker* incorporates depth-enhanced focused regions of the image into the thinking process, enabling **S**patially-aware **I**mage **F**ocus. *SIFThinker* can iteratively analyze and refine the regions of interest, ultimately delivering a more accurate final response. In the following sections, we provide a detailed description of the data generation pipeline, the spatially-aware image focus training paradigm, and the **GRPO-SIF**.

Data Generation

To simulate the way humans observe spatial scenes, we model a focus mechanism that incorporates depth for data generation. That is, we construct **SIF-50K**, a dataset consisting of two parts: (1) a tailored fine-grained reasoning

subset derived from spatial scenes in Flickr30k (Plummer et al. 2015), Visual7W (Zhu et al. 2016), GQA (Hudson and Manning 2019), Open Images (Kuznetsova et al. 2020), VSR (Liu, Emerson, and Collier 2023), and Birds-200-2021 (Wah et al. 2011), based on VisCoT (Shao et al. 2024b); and (2) a multi-instances subset resampled from TallyQA (Acharya, Kafle, and Kanan 2019). All source datasets include ground-truth bounding boxes (b-boxes) annotations.

As illustrated in Alg. 1, given each question-image-b-boxes-answer pairs (Q, I, B_{gt}, R), we apply a reverse expansion procedure, followed by forward reasoning over the expanded regions based on DepthAnythingV2 (Yang et al. 2024) and Doubao-1.5-vision-pro (Guo et al. 2025a). This process yields the final SIF-50K dataset, denoted as $\mathcal{P} = \{(Q, I, D_I, B_{gt}, R, R_{cot})\}$.

Spatially-aware Image Focus Training Paradigm

Method Overview. We propose a two-stage pipeline to incorporate spatially-aware grounded reasoning. The first stage is a warm-start supervised fine-tuning phase, which biases the model to generate structured reasoning chains with explicitly focused regions, resulting in $\mathcal{M}_{SFT}$. This is followed by a reinforcement learning phase that further refines and optimizes these grounded behaviors, yielding the final model $\mathcal{M}_{RL}$. For SFT, we utilize the full set from SIF-50K to get $\mathcal{P}_{SFT} = (Q, I, D_I, R_{cot})$. For RL (detailed in the following section), to promote progressive learning with minimal supervision, we sample 200 instances from SIF-50K to form a smaller set $\mathcal{P}_{RL} = (Q, I, D_I, B_{gt}, R)$.

Reinforcement Learning with GRPO-SIF

RL Formulation. Built upon the Group-Relative Policy Optimisation (GRPO) (Shao et al. 2024c), the model $\mathcal{M}_{SFT}$ is framed as a policy π_θ that generates an output sequence conditioned on the input (Q, I, D_I). During training, for each question-image-depthimage pair (Q, I, D_I), GRPO-SIF sample a set of N candidate completions $\{o_1, \ldots, o_N\}$ from the current policy $\pi_{\theta_{\text{old}}}$, and then updates the current policy π_θ by maximizing the following objective:

$$\mathcal{J}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Big\{\min\Big[\text{clip}(r_{i,t}, 1-\epsilon, 1+\epsilon)\hat{A}_{i,t}, r_{i,t}\hat{A}_{i,t}\Big] - \beta\,\mathbb{D}_{\text{KL}}[\pi_\theta \,||\, \pi_{\text{ref}}]\Big\},$$

where $r_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}$ denotes the ratio between the new and old policies at step t, ϵ and β are hyperparameters. $\mathbb{D}_{\text{KL}}[\pi_\theta \,||\, \pi_{\text{ref}}]$ estimate the KL divergence with the unbiased estimator (Schulman 2020) between current policy model and the reference model. For each completion o_i, a task-specific reward $r_{i,t} = R(Q, I, D_I, B_{gt}, R, o_i)$ is computed based on a combination of reward components (detailed below) at step t. These rewards are then used to compute a group-normalized advantage:

$$\hat{A}_{i,t} = \frac{r_{i,t} - \text{mean}\{r_{1,t}, \ldots, r_{N,t}\}}{\text{std}\{r_{i,t}, \ldots, r_{N,t}\} + \delta}. \tag{1}$$

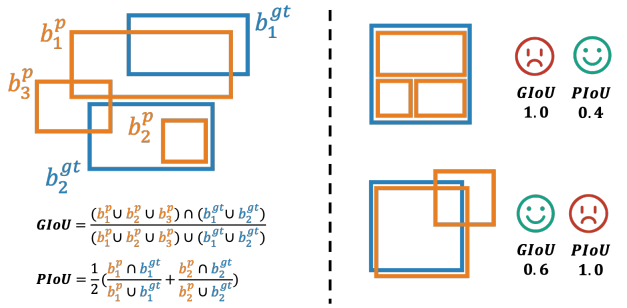

Figure 2: Visualization of our proposed $HIoU$ (left). The performance of $GIoU$ and $PIoU$ are illustrated respectively (right), highlighting the robustness against reward hacking.

δ is a small constant (e.g., 10^{-8}) added for numerical stability. The task reward $r_{i,t}$ is a composite signal comprising four components: a spatially-aware grounded reasoning-format reward (r_{format}), a progressive-answer-accuracy reward ($r_{\text{ans},t}$), a correction-enhanced-grounding reward (r_{bbox}), and depth-consistency reward (r_{depth}). These components are designed to jointly encourage spatially-aware grounded reasoning, thus attaching the precise answer.

Hierarchical Intersection over Union ($HIoU$). To comprehensively reward the grounding quality between predicted b-boxes $B_p = \{b_1^p, b_2^p, \ldots, b_n^p\}$ and ground-truth b-boxes $B_{gt} = \{b_1^{gt}, b_2^{gt}, \ldots, b_m^{gt}\}$, we propose the Hierarchical IoU ($HIoU$) as shown in Fig. 2. This design mitigates reward hacking issues such as inflated AP reward scores by incorporating both global and instance-level complementary components. (1) We first compute the global IoU ($GIoU$), which quantifies the overall spatial consistency between the predicted b-boxes and ground-truth b-boxes as a whole:

$$GIoU = \frac{\left|\bigcup_{b_i^p \in B_p} b_i^p \cap \bigcup_{b_j^{gt} \in B_{gt}} b_j^{gt}\right|}{\left|\bigcup_{b_i^p \in B_p} b_i^p \cup \bigcup_{b_j^{gt} \in B_{gt}} b_j^{gt}\right|}. \tag{2}$$

(2) Next, we compute a pairwise IoU via one-to-one bipartite matching between predicted b-boxes and ground-truth b-boxes using the Kuhn-Munkres algorithm (Kuhn 1955). Let $\mathcal{M}' \subseteq B_p \times B_{gt}$ denote the optimal matching set that maximizes the total IoU:

$$\mathcal{M} = \arg\max_{\mathcal{M}' \subseteq B_p \times B_{gt}} \sum_{(i,j)\in\mathcal{M}'} \text{IoU}(b_i^p, b_j^{gt}), \quad \text{s.t.} \quad \mathcal{M}' \text{ is one-to-one.} \tag{3}$$

The pairwise IoU($PIoU$) score is then computed as the average over the matched pairs:

$$PIoU = \frac{1}{|\mathcal{M}|}\sum_{(i,j)\in\mathcal{M}} \frac{|b_i^p \cap b_i^{gt}|}{|b_i^p \cup b_i^{gt}|}. \tag{4}$$

The final $HIoU$ is computed as the average of the global and pairwise IoU accuracy.

$$HIoU = \frac{GIoU + PIoU}{2}. \tag{5}$$

Reasoning-format Reward (r_{format}). This reward encourages the model to generate reasoning outputs structured with the designated special token, specifically adhering to the format: `<think><area> ... </area><text> ... </text></think><answer> ... </answer>`. The `<area>...</area>` must contain a JSON-formatted representation of bounding boxes with depth, while `<text>...</text>` provides an explanation grounded in the specified spatial region. A reward of 1.0 is assigned to responses that strictly comply with this format.

Progressive-answer-accuracy Reward ($r_{\text{ans},t}$). This reward integrates both the correctness of the final answer and the progression of answer quality over time, yielding a more robust signal than purely rule-based evaluations. Specifically, we leverage an external Vision-Language Model (*Doubao-1.5-vision-pro*) as the judge to assess response quality. The reward is defined as:

$$r_{\text{ans},t} = s_t + (s_t - \text{mean}\{s_{1,t-1}, \ldots, s_{N,t-1}\}) \quad (6)$$

where s_t denotes the continuous score assigned by the Doubao judge at step t, based on the question, the predicted answer, and the ground-truth answer. The term $(s_t - \text{mean}\{s_{1,t-1}, \ldots, s_{N,t-1}\})$ captures the improvement between consecutive steps, thereby encouraging progressive refinement of the model's responses.

Correction-enhanced grounding Reward (r_{bbox}). Given the structured nature of our output format, we can explicitly extract the sequence of bounding boxes generated during the reasoning process, thereby enabling fine-grained tracking of the step-by-step grounding. Let B_{ini} denote the first bounding boxes in the reasoning trajectory that does not cover the entire image (i.e., $[0, 0, 1, 1]$), and let B_{end} denote the final bounding boxes in the sequence. To assess the accuracy of the grounding results, we compute two intermediate metrics: (1) the $HIoU$ between the final box B_{end} and the ground-truth box B_{gt}, denoted as $s_{\text{end}} = HIoU(B_{\text{end}}, B_{gt})$; and (2) the $HIoU$ between the initial box B_{ini} and B_{gt}, denoted as $s_{\text{init}} = HIoU(B_{\text{ini}}, B_{gt})$. The overall grounding reward consists of two components: the final grounding accuracy s_{end} and a correction-aware improvement term, defined as the relative gain $s_{\text{end}} - s_{\text{init}}$:

$$r_{\text{bbox}} = s_{\text{end}} + (s_{\text{end}} - s_{\text{init}}). \quad (7)$$

Depth-consistency Reward (r_{depth}). A spatially-aware model is expected to accurately capture the depth value associated with each specified region. However, the presence of hallucinations may cause the autoregressively generated next-token to become inconsistent with the actual input depth image, resulting in spatial confusion. To address this, we perform step-wise verification on the depth tokens generated during the reasoning process. Specifically, for each b-boxes-depth pair (B_i, d_i), we extract the corresponding ground-truth depth d_i^{gt} from the depth map D_I based on B_i, and require the absolute error to be less than the threshold $\mathcal{T} = 0.1$. A reward is assigned only when all depth values throughout the reasoning trajectory meet this consistency criterion.

$$r_{\text{depth}} = \mathbb{I}\left(\forall i : \frac{|d_i - d_i^{\text{gt}}|}{d_i^{\text{gt}}} \leq \mathcal{T}\right), \quad (8)$$

Model	SpatialBench				SAT-Sta.	CV-Bench
	Pos.	Ext.	Cnt.	Size		
o3-2025-04-16	79.4	95.0	89.9	35.0	67.4	—
LLaVA-1.5-7B	44.1	45.0	82.8	30.0	49.8	51.7
LLaVA-NeXT-7B	47.1	75.0	84.0	20.0	54.1	62.7
SpatialBot-3B	50.0	80.0	86.7	25.0	61.5	65.1
Emu3-8B	47.1	20.0	10.0	25.0	—	—
Bunny-8B	50.0	75.0	90.4	**25.0**	60.8	61.0
SpatialBot-8B	53.0	75.0	90.4	20.0	—	—
SIFThinker-8B	**61.8**	**80.0**	**92.2**	23.3	**67.9**	**65.6**
Qwen2.5-VL-7B	61.8	80.0	87.1	30.0	65.5	73.0
SSR-7B	64.7	85.0	90.2	28.3	—	73.3
SIFThinker-7B	**73.5**	**95.0**	**94.7**	**35.0**	**72.8**	**75.9**

Table 1: **Spatial perception evaluation** on SpatialBench (Position, Existence, Counting, Size), SAT (Static), and CV-Bench. Bunny-LLaMA3-8B and Qwen2.5-VL-7B serve as base models in 3- and 4-th groups. The **best** is highlighted.

where $\mathbb{I}(\cdot)$ is the indicator function.

Experiments

We evaluate *SIFThinker* with several SOTA methods on an array of different categories as follows.

Spatial Intelligence

We evaluate our method against several SOTA methods on a range of spatial understanding benchmarks. Benefiting from our spatially-aware think-with-images training paradigm, our model demonstrates superior 3D understanding capabilities. As shown in Tab. 1, under the same base model, our method outperforms SpatialBot (Cai et al. 2024) by **7.82%** (64.3 vs. 59.6) and SSR (Liu et al. 2025b) by **11.17%** (74.5 vs. 67.1) on SpatialBench (Cai et al. 2024). Furthermore, we evaluate our method on larger-scale benchmarks, SAT (Static) (Ray et al. 2024) and CV-Bench (Tong et al. 2024), achieving gains of **11.15%** (72.8 vs. 65.5) and **3.97%** (75.9 vs. 73.0) over the Qwen2.5-VL-7B base model, respectively. Although both SpatialBot and SSR incorporate depth images to enhance spatial understanding, *we argue that depth perception and spatial grounding are inherently complementary*. By introducing reasoning over spatially grounded regions, our method achieves more significant improvements.

We further conducted a comparison with the representative SOTA closed-source model—ChatGPT-o3 (OpenAI 2025). On SpatialBench, *SIFThinker* achieves a comparable average score to o3 (74.6 vs. 74.8). Notably, on SAT-Static, our method even outperforms o3 by a significant margin of **8.01%** (72.8 vs. 67.4), demonstrating the superior capability of *SIFThinker* in spatial perception.

Visual Perception

In this section, we comprehensively evaluate the visual perception capabilities of the method in terms of visual understanding, grounding capability and self-correction ability.

Visual Understanding. We select scene-related (e.g. non-planar) subsets from the VisCoT as VisCoT_s, and choose the attribute and spatial subsets from V*Bench. As shown

Model	VisCoT_s						V*Bench	
	Flickr30k	GQA	Open images	VSR	CUB	Avg	Attribute	Spatial
o3-2025-04-16 (OpenAI 2025)	0.828	0.706	0.515	0.826	0.933	0.762	0.739	0.803
LLaVA-1.5-7B (Liu et al. 2023)	0.581	0.534	0.412	0.572	0.530	0.526	0.435	0.566
VisCoT-LLaVA-1.5-7B (Shao et al. 2024b)	0.671	0.616	**0.833**	0.682	0.556	0.672	0.466	0.571
SIFThinker-LLaVA-1.5-7B	**0.749**	**0.675**	0.801	**0.698**	**0.831**	**0.751**	**0.565**	**0.671**
Qwen2.5-VL-7B (Bai et al. 2025a)	0.601	0.467	0.289	0.581	0.583	0.504	0.644	0.634
VisRL-Qwen2.5-VL-7B# (Chen, Luo, and Li 2025)	0.662	0.589	0.767	0.698	0.772	0.698	0.678	0.658
SEAL-7B (V*) (Wu and Xie 2024)	0.723	0.599	0.448	0.730	0.640	0.628	0.748	0.763
SIFThinker-Qwen2.5-VL-7B	**0.755**	**0.664**	**0.773**	**0.734**	**0.872**	**0.760**	**0.791**	**0.776**

Table 2: **Visual perception performance** on VisCoT_s and V*Bench. # indicates methods trained on the same **SIF-50K** dataset as ours. For the same base models, the **best** is highlighted.

Model	RefCOCO			RefCOCO+			RefCOCOg		OVDEval
	val	test-A	test-B	val	test-A	test-B	val-u	test-u	
UNIEXT (Yan et al. 2023)	92.6	94.3	**91.5**	85.2	89.6	79.8	88.7	89.4	—
Grounding-DINO (Liu et al. 2024a)	90.6	93.2	88.2	82.8	89.0	75.9	86.1	87.0	25.3
Qwen2.5-VL-3B (Bai et al. 2025a)	89.1	91.7	84.0	82.4	88.0	74.1	85.2	85.7	25.5
Qwen2.5-VL-7B (Bai et al. 2025a)	90.0	92.5	85.4	84.2	89.1	76.9	87.2	87.2	29.1
VLM-R1-Qwen2.5-VL-3B (Shen et al. 2025)	91.2	92.9	87.3	84.8	88.1	76.8	87.8	87.9	31.0
VLM-R1-Qwen2.5-VL-7B# (Shen et al. 2025)	90.9	93.4	87.5	85.0	88.6	77.2	88.2	88.3	26.3
VisRL-Qwen2.5-VL-7B# (Chen, Luo, and Li 2025)	92.3	94.2	88.9	84.4	89.5	77.8	88.5	89.1	25.9
SIFThinker-Qwen2.5-VL-7B	**93.8**	**95.1**	90.4	**86.0**	**90.7**	**80.5**	**90.4**	**90.6**	**37.8**

Table 3: Performance (Top-1 Accuracy@0.5) on Referring Expression Comprehension tasks and performance (NMS-AP) on Open-Vocabulary Detection tasks. # indicates methods trained on the same **SIF-50K** dataset as ours. The **best** is highlighted.

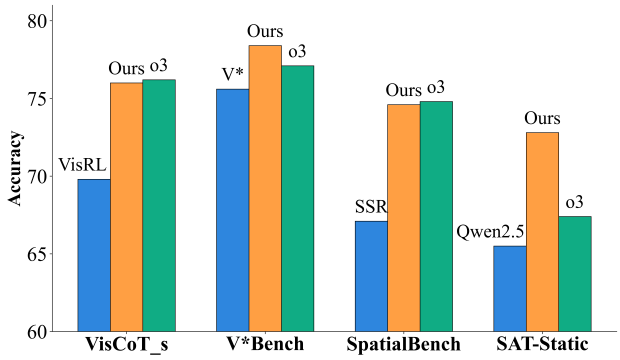

Figure 3: Comparison with other open-source SOTA methods (blue) under various benchmarks in terms of the same base model. Besides, we also include the performance of the proprietary SOTA model ChatGPT-o3-2025-04-16 (green).

in Tab. 2, on VisCoT_s dataset, under the same base model – LLaVA-1.5-7B, *SIFThinker* outperforms VisCoT by **11.76%** (0.751 vs. 0.672). Taking Qwen2.5-VL-7B as the base model for training, we surpass VisRL by **8.89%** (0.760 vs. 0.698). V*Bench presents a more challenging evaluation of fine-grained perception on high-resolution images. Remarkably, our method outperforms the state-of-the-art method – SEAL by **5.75%** on the Attribute subset (0.791 vs. 0.748), and by **1.70%** on the Spatial subset (0.776 vs. 0.763). In contrast to VisCoT, VisRL and SEAL, *SIFThinker does not rely on a staged crop-image process*.

ChatGPT-o3 leverages an external image magnification tool to explicitly capture fine-grained regions within the image. In contrast, *SIFThinker* adopts an intrinsic integration of visual information without any external tools. As shown in Tab. 2, a detailed comparison reveals that our method achieves comparable overall performance to o3 on the VisCoT_s benchmarks (0.760 vs. 0.762). Furthermore, on V*Bench-Attribute, our method surpasses o3 by a substantial margin of **7.04%** (0.791 vs. 0.739), highlighting the effectiveness in fine-grained visual reasoning. Instead, *SIFThinker* integrates visual perception directly into the thinking process with dynamically evolving bounding boxes to represent shifts in attention. This results in a more unified and intrinsically coherent reasoning approach. Furthermore, we explicitly incorporate 3D spatial information, which brings additional benefits to scene-level grounding.

Grounding Capability. Since the accuracy of the final response is partly dependent on the correctness of the bounding boxes generated during the think process, we further evaluate the model's grounding capability as shown in Tab. 3. Specifically, we select two structurally similar tasks—Referring Expression Comprehension (REC) and Open-Vocabulary Detection (OVD)—both of which require the model to generate bounding boxes conditioned on textual descriptions. *SIFThinker* outperforms all previous generalist models with comparable parameters, achieving an average improvement of **1%** to **3%** over VisRL (93.08 vs. 91.78 on RefCOCO (Kazemzadeh et al. 2014), 85.76 vs. 83.90 on RefCOCO+ (Mao et al. 2016), 90.47 vs. 88.82 on RefCOCOg (Mao et al. 2016)). Moreover, in most of cases, *SIFThinker* even surpasses previous state-of-the-art specialist models (e.g. Grounding-DINO, UNINEXT). To further

Model	MME^P	MME^C	MMB^T	MMB^D	SEED-I	VQA^{v2}	POPE
LLaVA	1511	282	66.5	62.1	65.8	79.1	85.9
VisCoT	1454	**308**	69.2	**66.6**	66.0	81.3	86.0
SIFThinker	**1531**	295	**69.3**	62.8	**66.0**	**81.6**	**86.6**
Bunny	1574	342	73.7	74.2	72.3	80.5	85.2
SpatialBot	1576	333	75.8	74.8	72.4	80.9	85.3
SIFThinker	**1580**	**355**	**76.8**	**75.0**	**72.5**	**81.8**	**85.7**
Qwen2.5VL	1670	623	80.3	81.5	77.0	83.3	85.9
SIFThinker	**1702**	**650**	**83.4**	**82.4**	**77.2**	**84.5**	**86.9**

Table 4: Results on general VLM Benchmarks. Methods are grouped by their underlying base model: LLaVA-1.5-7B (top), Bunny-Llama3-8B (middle), and Qwen2.5-VL-7B (bottom). For each group, the **best** is highlighted.

assess multi-object grounding performance of our method, we adopt OVDEval (Yao et al. 2023) with NMS-AP as the evaluation metric. Empowered by our proposed $HIoU$, *SIFThinker* exhibits promising robustness in multi-object scenarios, outperforming VisRL by **45.9%** (37.8 vs. 25.9).

General VLM Benchmarks

As shown in Tab. 4, we report results on widely used general benchmarks, including MME (Fu et al. 2024) perception (MME^P), MME cognition (MME^C), MMBench (Liu et al. 2024b) test and dev sets (denoted as MMB^T and MMB^D), SEED-Bench (Li et al. 2023a) images (SEED-I), VQA^{v2} (Goyal et al. 2017) test-dev split, and POPE (Li et al. 2023c) (measured as the average F1 score over three categories on the COCO validation set). Across most of these benchmarks, *SIFThinker* not only avoids performance degradation but even achieves notable improvements, demonstrating the robustness of our method—particularly in scenarios that depth information can benefit. Under the same base models, *SIFThinker* consistently outperforms both VisCoT, which focuses on fine-grained visual perception, and SpatialBot, which emphasizes spatial reasoning. Remarkably, on MMB^T, *SIFThinker* achieves about **4%** improvements across different base model settings (69.3 vs. 66.5 on LLaVA-1.5-7B, 76.8 vs. 73.7 on Bunny-Llama3-8B, 83.4 vs. 80.3 on Qwen2.5-VL-7B).

Overall, as shown in Fig. 3, at the 7B parameter scale, *SIFThinker* (orange) exhibits strong capabilities in both spatial understanding and visual perception, outperforming all open-source models (blue) and approaching the performance of the larger-scale ChatGPT-o3 (green). Notably, it even surpasses ChatGPT-o3 on V*Bench and SAT-Static.

Ablation Study

Training Strategy. We present a comprehensive ablation study in Tab. 5. VQA-SFT refers to directly applying SFT on the original question-answer pairs, which serve as the source data from where **SIF-50K** was constructed, whereas CoT-SFT leverages the Chain-of-Thought (CoT) construction strategy introduced in Alg. 1. This demonstrates that guiding the model to think with images yields a notable performance improvement of **8.58%** (0.582 vs. 0.536). However, SFT alone mainly helps the model learn output formatting, and in some cases (e.g., GQA), even leads to performance degradation. In contrast, incorporating RL yields consistent and substantial improvements, achieving an additional **30.58%** gain over SFT alone (0.760 vs. 0.582).

Model	VisCoT_s					
	Flickr30k	GQA	Open images	VSR	CUB	Avg
VQA-SFT	0.542	0.508	0.390	0.597	0.642	0.536
CoT-SFT	0.517	0.468	0.556	0.588	0.780	0.582
w/o r_{bbox}	0.641	0.580	0.711	0.705	0.839	0.695
w/o $r_{\text{ans},t}$	0.628	0.573	0.658	0.663	0.795	0.663
w/o D_I	0.571	0.595	0.589	0.651	0.870	0.655
w/o r_{depth}	0.748	**0.666**	0.762	0.718	**0.872**	0.753
Full	**0.755**	0.664	**0.773**	**0.734**	**0.872**	**0.760**

Table 5: Ablated settings in terms of Qwen2.5-VL-7B.

Component Removal. We further perform ablations on various RL rewards (w/o $r_{\text{ans},t}$, r_{bbox}, r_{depth}) and evaluate the impact of depth information (w/o D_I) in Tab. 5. The results suggest that the observed performance gains are primarily attributed to three key factors: 1) the think-with-images reasoning paradigm, which promotes spatially grounded cognition; 2) the carefully designed reward functions for both bounding box prediction and response generation, which work synergistically to encourage iterative correction and refinement; 3) the inclusion of depth inputs, which enhance the model's spatial intelligence during grounding. Together, these designs form a unified and robust framework for spatially-aware visual grounding, endowing the model with a general-purpose reasoning capability that improves performance across diverse benchmarks.

Effectiveness with RGB-Depth Input. When both RGB and depth are provided, *SIFThinker* still outperforms baselines—e.g., Qwen2.5-VL: ours vs. VisRL = **0.760** vs. 0.691; LLaVA-1.5: ours vs. VisCoT = **0.751** vs. 0.674.

Ablation on Correction-enhanced Grounding Reward. Removing $(s_{end} - s_{init})$ reduces performance on VisCoT_s from **0.760** to 0.723. We also experimented with a step-wise non-decreasing-IoU reward, which also drops from **0.760** to 0.735 due to restricted exploration.

Conclusion and Limitation

In this paper, we propose *SIFThinker*, a spatially-aware image-text interleaved reasoning framework. Inspired by human-like prompt-driven search in 3D environments, *SIFThinker* performs spatially-aware grounding before delivering the final response. Specifically, we introduce a novel pipeline for generating CoT datasets tailored for think-with-images reasoning, enabling process-level supervision. Building on this dataset, we propose GRPO-SIF by incorporating not only region-level corrective signals, but also proposing progress learning and depth consistency rewards. Extensive experiments across diverse benchmarks demonstrate the effectiveness of *SIFThinker*.

Limitation & Future Work. Since *SIFThinker* is trained on single-image, it may face challenges in generalizing to dynamic spatial scenarios that require reasoning across multiple images. We believe extending it to such settings would be interesting for higher practical impact.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under contract No. 62171256.

References

Acharya, M.; Kafle, K.; and Kanan, C. 2019. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8076–8084.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025a. Qwen2. 5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Bai, S.; Li, M.; Liu, Y.; Tang, J.; Zhang, H.; Sun, L.; Chu, X.; and Tang, Y. 2025b. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*.

Cai, W.; Ponomarenko, I.; Yuan, J.; Li, X.; Yang, W.; Dong, H.; and Zhao, B. 2024. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*.

Chen, B.; Xu, Z.; Kirmani, S.; Ichter, B.; Sadigh, D.; Guibas, L.; and Xia, F. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14455–14465.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2024b. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, 370–387. Springer.

Chen, L.; Li, L.; Zhao, H.; Song, Y.; and Vinci. 2025a. R1-V: Reinforcing Super Generalization Ability in Vision-Language Models with Less Than $3. https://github.com/Deep-Agent/R1-V. Accessed: 2025-02-02.

Chen, L.; Wei, X.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Tang, Z.; Yuan, L.; et al. 2025b. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37: 19472–19495.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A. J.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Chen, Z.; Luo, X.; and Li, D. 2025. Visrl: Intention-driven visual perception via reinforced reasoning. *arXiv preprint arXiv:2503.07523*.

Cheng, A.-C.; Yin, H.; Fu, Y.; Guo, Q.; Yang, R.; Kautz, J.; Wang, X.; and Liu, S. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37: 135062–135093.

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; Huang, W.; et al. 2023. Palm-e: An embodied multimodal language model.

Fan, Y.; He, X.; Yang, D.; Zheng, K.; Kuo, C.-C.; Zheng, Y.; Narayanaraju, S. J.; Guan, X.; and Wang, X. E. 2025. GRIT: Teaching MLLMs to Think with Images. *arXiv preprint arXiv:2505.15879*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; and Ji, R. 2024. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. arXiv:2306.13394.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6904–6913.

Guo, D.; Wu, F.; Zhu, F.; Leng, F.; Shi, G.; Chen, H.; Fan, H.; Wang, J.; Jiang, J.; Wang, J.; et al. 2025a. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025b. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; and Lew, M. S. 2016. Deep learning for visual understanding: A review. *Neurocomputing*, 187: 27–48.

Hu, J.; Wu, X.; Zhu, Z.; Xianyu; Wang, W.; Zhang, D.; and Cao, Y. 2024. OpenRLHF: An Easy-to-use, Scalable and High-performance RLHF Framework. *arXiv preprint arXiv:2405.11143*.

Hudson, D. A.; and Manning, C. D. 2019. 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*, 6700–6709.

Ji, J.; Zhou, J.; Lou, H.; Chen, B.; Hong, D.; Wang, X.; Chen, W.; Wang, K.; Pan, R.; Li, J.; Wang, M.; Dai, J.; Qiu, T.; Xu, H.; Li, D.; Chen, W.; Song, J.; Zheng, B.; and Yang, Y. 2024. Align Anything: Training All-Modality Models to Follow Instructions with Language Feedback.

Jiao, Y.; Chen, S.; Jie, Z.; Chen, J.; Ma, L.; and Jiang, Y.-G. 2025. Lumen: Unleashing versatile vision-centric capabilities of large multimodal models. *Advances in Neural Information Processing Systems*, 37: 81461–81488.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 787–798.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7): 1956–1981.

Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; and Shan, Y. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Li, C.; Wu, W.; Zhang, H.; Xia, Y.; Mao, S.; Dong, L.; Vulić, I.; and Wei, F. 2025. Imagine while Reasoning in Space: Multimodal Visualization-of-Thought. *arXiv preprint arXiv:2501.07542*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Liu, F.; Emerson, G.; and Collier, N. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11: 635–651.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023. Improved Baselines with Visual Instruction Tuning.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024a. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 38–55. Springer.

Liu, Y.; Chi, D.; Wu, S.; Zhang, Z.; Hu, Y.; Zhang, L.; Zhang, Y.; Wu, S.; Cao, T.; Huang, G.; et al. 2025a. SpatialCoT: Advancing Spatial Reasoning through Coordinate Alignment and Chain-of-Thought for Embodied Task Planning. *arXiv preprint arXiv:2501.10074*.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, 216–233. Springer.

Liu, Y.; Ma, M.; Yu, X.; Ding, P.; Zhao, H.; Sun, M.; Huang, S.; and Wang, D. 2025b. SSR: Enhancing Depth Perception in Vision-Language Models via Rationale-Guided Spatial Reasoning. *arXiv preprint arXiv:2505.12448*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.

OpenAI. 2025. Introducing OpenAI o3 and o4-mini. Https://openai.com/index/introducing-o3-and-o4-mini/.

Palmeri, T. J.; and Gauthier, I. 2004. Visual object understanding. *Nature Reviews Neuroscience*, 5(4): 291–303.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, 2641–2649.

Qi, J.; Ding, M.; Wang, W.; Bai, Y.; Lv, Q.; Hong, W.; Xu, B.; Hou, L.; Li, J.; Dong, Y.; et al. 2024. Cogcom: Train large vision-language models diving into details through chain of manipulations.

Ray, A.; Duan, J.; Tan, R.; Bashkirova, D.; Hendrix, R.; Ehsani, K.; Kembhavi, A.; Plummer, B. A.; Krishna, R.; Zeng, K.-H.; et al. 2024. SAT: Spatial Aptitude Training for Multimodal Language Models. *arXiv preprint arXiv:2412.07755*.

Schulman, J. 2020. Approximating kl divergence. *John Schulman's Homepage*.

Shao, H.; Hu, Y.; Wang, L.; Song, G.; Waslander, S. L.; Liu, Y.; and Li, H. 2024a. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15120–15130.

Shao, H.; Qian, S.; Xiao, H.; Song, G.; Zong, Z.; Wang, L.; Liu, Y.; and Li, H. 2024b. Visual CoT: Unleashing Chain-of-Thought Reasoning in Multi-Modal Language Models. arXiv:2403.16999.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024c. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shen, H.; Zhang, Z.; Zhang, Q.; Xu, R.; and Zhao, T. 2025. VLM-R1: A stable and generalizable R1-style Large Vision-Language Model. https://github.com/om-ai-lab/VLM-R1. Accessed: 2025-02-15.

Su, Z.; Li, L.; Song, M.; Hao, Y.; Yang, Z.; Zhang, J.; Chen, G.; Gu, J.; Li, J.; Qu, X.; et al. 2025. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*.

Thawakar, O.; Dissanayake, D.; More, K.; Thawkar, R.; Heakl, A.; Ahsan, N.; Li, Y.; Zumri, M.; Lahoud, J.; Anwer, R. M.; Cholakkal, H.; Laptev, I.; Shah, M.; Khan, F. S.; and Khan, S. 2025. LlamaV-o1: Rethinking Step-by-step Visual Reasoning in LLMs. arXiv:2501.06186.

Tong, P.; Brown, E.; Wu, P.; Woo, S.; IYER, A. J. V.; Akula, S. C.; Yang, S.; Yang, J.; Middepogu, M.; Wang, Z.; et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37: 87310–87356.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, X.; Zhang, S.; Li, S.; Kallidromitis, K.; Li, K.; Kato, Y.; Kozuka, K.; and Darrell, T. 2024. SegLLM: Multi-round Reasoning Segmentation. *arXiv preprint arXiv:2410.18923*.

Wolfe, J. M.; and Horowitz, T. S. 2017. Five factors that guide attention in visual search. *Nature human behaviour*, 1(3): 0058.

Wu, M.; Yang, J.; Jiang, J.; Li, M.; Yan, K.; Yu, H.; Zhang, M.; Zhai, C.; and Nahrstedt, K. 2025. VTool-R1: VLMs Learn to Think with Images via Reinforcement Learning on Multimodal Tool Use. *arXiv preprint arXiv:2505.19255*.

Wu, P.; and Xie, S. 2024. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13084–13094.

Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T.-S. 2024. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.

Yan, B.; Jiang, Y.; Wu, J.; Wang, D.; Luo, P.; Yuan, Z.; and Lu, H. 2023. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15325–15336.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911.

Yao, Y.; Liu, P.; Zhao, T.; Zhang, Q.; Liao, J.; Fang, C.; Lee, K.; and Wang, Q. 2023. How to Evaluate the Generalization of Detection? A Benchmark for Comprehensive Open-Vocabulary Detection. *arXiv preprint arXiv:2308.13177*.

Zhang, J.; Jiao, Y.; Chen, S.; Chen, J.; and Jiang, Y.-G. 2024. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*.

Zhang, X.; Gao, Z.; Zhang, B.; Li, P.; Zhang, X.; Liu, Y.; Yuan, T.; Wu, Y.; Jia, Y.; Zhu, S.-C.; et al. 2025. Chain-of-Focus: Adaptive Visual Search and Zooming for Multimodal Reasoning via RL. *arXiv preprint arXiv:2505.15436*.

Zheng, Z.; Yang, M.; Hong, J.; Zhao, C.; Xu, G.; Yang, L.; Shen, C.; and Yu, X. 2025. DeepEyes: Incentivizing" Thinking with Images" via Reinforcement Learning. *arXiv preprint arXiv:2505.14362*.

Zhu, Y.; Groth, O.; Bernstein, M.; and Fei-Fei, L. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4995–5004.