Spatial Preference Rewarding for MLLMs Spatial Understanding

Han Qiu[1], Peng Gao[2], Lewei Lu[3], Xiaoqin Zhang[4], Ling Shao[5], Shijian Lu[1*]

[1]S-Lab, Nanyang Technological University, [2]Shanghai AI Laboratory
[3]Sensetime Research, [4]Zhejiang University of Technology
[5]UCAS-Terminus AI Lab,University of Chinese Academy of Sciences

han023@e.ntu.edu.sg, Shijian.Lu@ntu.edu.sg

Abstract

Multimodal large language models (MLLMs) have demonstrated promising spatial understanding capabilities, such as referencing and grounding object descriptions. Despite their successes, MLLMs still fall short in fine-grained spatial perception abilities, such as generating detailed region descriptions or accurately localizing objects. Additionally, they often fail to respond to the user's requirements for desired fine-grained spatial understanding. This issue might arise because existing approaches primarily focus on tuning MLLMs to model pre-annotated instruction data to inject spatial knowledge, without direct supervision of MLLMs' actual responses. We address this issue by SPR, a Spatial Preference Rewarding (SPR) approach that enhances MLLMs' spatial capabilities by rewarding MLLMs' detailed responses with precise object localization over vague or inaccurate responses. With randomly selected image regions and region descriptions from MLLMs, SPR introduces semantic and localization scores to comprehensively evaluate the text quality and localization quality in MLLM-generated descriptions. We also refine the MLLM descriptions with better localization accuracy and pair the best-scored refinement with the initial descriptions of the lowest score for direct preference optimization, thereby enhancing fine-grained alignment with visual input. Extensive experiments over standard referring and grounding benchmarks show that SPR improves MLLM spatial understanding capabilities effectively with minimal overhead in training. Data and code will be released at https://github.com/hanqiu-hq/SPR.

1. Introduction

Multimodal large language models (MLLMs) [3, 15, 18, 31, 32, 51, 58, 70] have achieved remarkable success by integrating pretrained large language model [2, 14, 49] with vision encoders [8, 37, 42], leading to significant advancements in a wide range of general vision-language tasks. By combining visual and language signals, MLLMs have demonstrated superior capabilities in multimodal understanding, reasoning, and interaction as compared with traditional vision models. Recently, several studies have further injected spatial knowledge into MLLMs, thereby improving MLLMs' fine-grained perception of visual inputs and enabling tasks such as referential dialogue [10, 16], grounding captioning [35, 58, 62], region description [31, 52], and object detection [61], etc. These advances have paved the way for MLLMs to serve as versatile visual assistants supporting a wider range of applications.

Despite recent advancements, MLLMs still face challenges in fine-grained spatial understanding, with responses not aligned with human preferences. As illustrated in 1, the generated grounded region descriptions are often vague with inaccurate object localizations, and models may fail to focus on the queried region, distracted from other regions in the image. The issue in spatial understanding could be attributed to the lack of positive-negative preference feedback in existing instruction-tuned MLLMs. Specifically, instruction fine-tuning (SFT) directly optimizes MLLMs to mimic ground truth positive samples, but it cannot impose any penalties if the model produces inaccurate negative samples for localization during actual inference. As a result, MLLMs may struggle to generate positive descriptions with accurate object localization and instead produce negative and inaccurate descriptions, leading to responses that do not align with user expectations. In addition, optimization using positive and negative samples has been proven crucial for spatial understanding in traditional object detection algorithms [7, 29, 45], highlighting a significant gap in the current MLLM training on spatial understanding.

Several studies [38, 48, 50, 59, 65, 69] attempt to introduce preference optimization for better MLLM alignment, where the preference data are constructed by collect-

*Corresponding author.

Figure 1. The proposed Spatial Preference Rewarding (SPR) mitigates the distracted and inaccurate region descriptions generated by MLLMs. Given an image and a user-specified region of interest, MLLMs often fail to focus on the queried region. They may be distracted by objects outside the specified region, failing to ground the queried objects, or providing inaccurate localization. Tuning MLLMs with our proposed SPR leads to more accurate object localization and detailed object descriptions.

ing MLLM-generated image descriptions and scoring them by human or LLMs. However, these methods primarily leverage preferences to improve image-level coarse alignment, and most of them target mitigating hallucinations in MLLMs. The problem of fine-grained alignment for spatial understanding, such as detailed region descriptions and accurate object localization, has been largely neglected.

To address this gap, we design SPR, a **S**patial **P**reference **R**ewarding framework that enhances MLLM spatial understanding capabilities by rewarding detailed responses with accurate object localization over vague or inaccurate responses. Specifically, SPR selects random image regions containing multiple objects and prompts MLLMs in diverse ways to generate grounded region descriptions. In reward modeling, it introduces both semantic and localization scores to evaluate the alignment between the region description and the region semantics, as well as how detailed region objects are described. We also refine the grounded object in the generated description to enhance its localization accuracy. Finally, the best-scored refined description and the response of the lowest score are paired as preferred and rejected data for direct preference optimization (DPO) [43] training with LORA [17]. By aligning MLLMs with detailed and accurate responses, SPR mitigates MLLMs' incompetence in accurate localization and spatial understanding as required in many real-world tasks.

We validate the effectiveness of SPR in enhancing MLLMs' spatial understanding capabilities with minimal overhead in training. Compared to the baseline, SPR enhances MLLMs on both referring and grounding benchmarks, especially under higher IoU thresholds which demand higher localization accuracy. In addition, SPR can improve MLLM trustworthiness and reduce MLLM hallucinations as well. Our experiments highlight the importance of incorporating preference-based feedback to enhance the fine-grained spatial understanding abilities in MLLMs.

The contributions of this work are summarized as follows:

- We propose a Spatial Preference Rewarding (SPR) framework to enhance the fine-grained spatial understanding of MLLMs via direct preference optimization (DPO), enhancing MLLMs' capabilities in precise region referring and accurate object localization in images.
- We develop an automated pipeline that creates preference data by constructing random region prompts and scoring model responses for spatial understanding. The pipeline requires no other MLLMs or human labours, making it scalable in future training.
- Extensive experiments show that the proposed SPR improves MLLMs' spatial understanding capabilities consistently across multiple public benchmarks.

2. Related Work

Multi-Modal Large Language Models (MLLMs.). Recently, the success of large language models (LLMs) [14, 49, 51] has been extended into the multimodal domain, resulting in models that demonstrate impressive performance in integrating vision and language [1, 15, 20, 32, 70]. These models treat visual signals as a special form of language, establishing multimodal understanding, reasoning, and interaction capabilities by combining visual encoders [37, 42] with pre-trained large language models, or by directly feeding encoded visual signals into LLMs [6, 53]. Most current MLLMs follow a two-step training process. The first step is pre-training, where large-scale vision-language datasets [9] are used to align visual features to the same space as language features. This enables the model to bridge visual and language embeddings effectively. The second step involves instruction-following finetuning, where high-quality vision-language datasets [25, 33, 70] are used to further enhance the MLLMs' capabilities to follow user instructions and comprehend multimodal information. These methods often convert existing datasets into an instruction-following format or adopt leading MLLMs like GPT to generate high-quality training instruction data for MLLMs [11, 33]. Despite their success, current MLLMs still face challenges that may generate undesired responses toward human preferences. For instance, these models are prone to generating hallucinated content [4, 30, 56, 59] or providing responses that do not fully meet user expectations. Improving the quality of MLLM responses and aligning them more closely with user preferences has thus become a surging focus of research in the community. Our work aims to improve the spatial understanding capabilities of MLLMs, aligning their behaviors better with human preferences.

MLLMs for Spatial Understanding. Spatial understanding capabilities [7, 13, 22, 67], such as object detection, referring, and grounding description tasks, have long been a fundamental research topic in the field of computer vision. Recent efforts attempt to empower MLLMs with dense visual perception and spatial understanding abilities by integrating region-level data in MLLM training or modifying MLLM architectures. For example, Kosmos-2 [39] and Shikra [10] directly represent the object coordinates in text, constructing instruction datasets to inject spatial knowledge into MLLMs. LLava-Grounding [63] and GroundingGPT [27] construct large-scale grounding datasets to enhance multimodal grounding capabilities. To better facilitate localization within images, RegionGPT [16], GPT4ROI [66], Ferret [58], and Groma [35] encode region features as direct inputs to LLMs, facilitating explicit attention to specific image regions. The Griffon [60, 61] series focuses on dense detection, enabling MLLMs to achieve performance comparable to traditional object detectors. LocVLM [44] explores the. LocVLM [44] explores key factors in instruction tuning for spatial understanding, such as coordinate representation, which improves MLLM's spatial awareness. However, these efforts primarily concentrate on the instruction-tuning phase and lack direct feedback on MLLMs' responses. To fill this gap, we propose a Spatial Preference Rewarding (SPR) framework, which constructs preference data based on MLLMs generated grounded region descriptions for MLLM tuning.

Preference Optimization for MLLMs. Preference alignment has recently emerged as a promising direction to align model responses with human preferences. One widely explored approach is to employ Reinforcement Learning from Human Feedback (RLHF) or Direct Preference Optimization (DPO) to improve the trustworthiness of MLLMs and reduce hallucinations in their responses. For example, LLaVA-RLHF [48] and RLHF-V [59] leverage human annotators to evaluate model responses and construct preference data for fine-tuning. POVID [68] and Silkie [26] use external models, such as GPT, as evaluators to build preference datasets. CLIP-DPO [38] and CSR [69] use CLIP to rank model responses to avoid resource-intensive human or MLLM annotations. AMP [65] introduced a multi-level preference framework to enable MLLMs to better model differences between preference data. mDPO [50] introduced additional preference data pairs with corrupted images to avoid over-optimization on language-only preferences. Unlike these existing studies that primarily aim to reduce hallucinations in MLLMs, our proposed SPR framework focuses on optimizing MLLM responses related to spatial reasoning and understanding. Specifically, SPR focuses on fine-grained alignment with visual inputs and facilitates MLLMs in distinguishing between high-quality object localization (positive samples) and inaccurate localization (negative samples), thereby improving the spatial understanding capabilities of MLLMs.

3. Methods

This section presents our proposed Spatial Preference Rewarding (SPR) framework. Following a typical DPO pipeline, SPR adopts a three-step process in MLLM finetuning, including collecting MLLMs' raw responses (Sec.3.1), evaluating the raw responses to construct preference data (Sec.3.2), and preference optimization (Sec.3.3). The details are elaborated in the following subsections.

3.1. Grounded Region Description Generation

The first step of our pipeline is to collect diverse model responses that will later be ranked to construct preference data. Since our primary objective is to enhance MLLMs' localization capabilities and achieve fine-grained alignment to visual inputs, we choose the task of region description with grounding to evaluate MLLMs' object localization capabilities. However, existing datasets [23, 57] for region

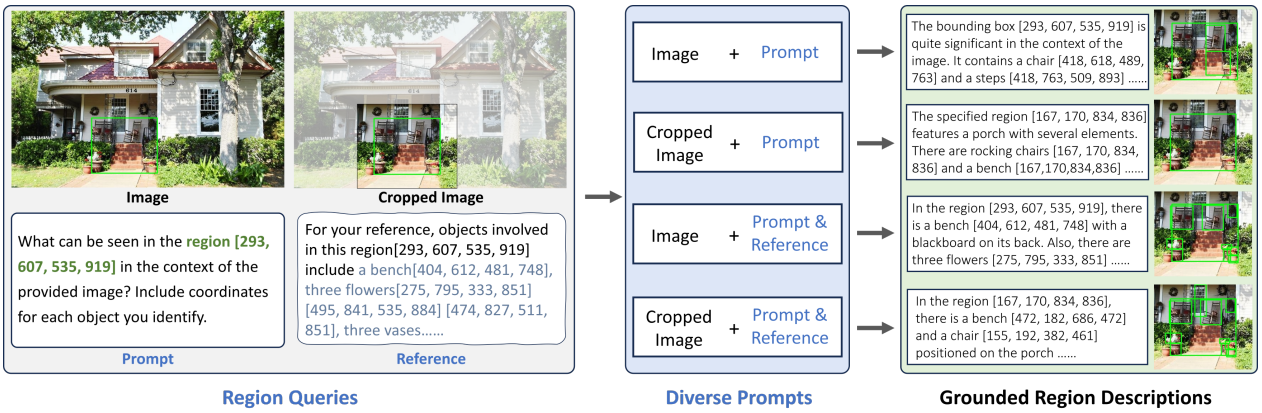

Figure 2. We leverage the generated object references and cropped image region to build a variety of multimodal prompts to enhance the diversity of generated region descriptions.

descriptions are often too simple, involving queried regions with only one or two objects and short phrases such as 'vehicles parked on the street' or 'bicycles are parked on the sidewalk.' Such simple prompts are inadequate for generating diverse responses to construct preferred and rejected preference data with sufficient divergence, which might hinder the effectiveness of DPO training [54]. To address this issue, we generate queried regions from scratch instead of using existing region description datasets.

Region Query Construction. We design a simple approach to generate randomly queried regions based on images and object annotations. Take the Objects365 dataset as an example. We first filter out images with few objects, ensuring that the data contains rich visual content. Then, given the annotated object bounding boxes in each image, we randomly select one of the objects as the starting region. From there, we iteratively expand the region by incorporating the nearest objects. The expansion stops randomly once more than four objects are involved in the region. The resulting region then prompts MLLMs to generate a detailed region description. Through this process, we simulate the human-like, dynamic attention across different parts of an image, encouraging the MLLM to adaptively focus on arbitrary image regions based on the given prompts.

Grounded Region Description Generation. As shown in Fig. 2, we build a variety of prompts for MLLMs to generate several region descriptions for each image, serving as candidate responses for preference data. Since the original MLLM sometimes struggles to generate detailed responses following region prompts, we utilize cropped region images along with object references constructed from annotations to guide MLLMs to attend to the region's content and details. These prompts help the model focus more effectively on the specified region and produce detailed descriptions that might better align with human preference. In this way, we encourage the MLLM to generate responses that are distinct in content but consistent in language style, which is then used for constructing preference data.

3.2. Preference Data Ranking and Construction

The next step is to rank the generated descriptions to obtain preferred and rejected data pairs. An ideal region description should meet at least two key criteria: (1) the text description should accurately match the semantics of the queried region and the surrounding image content, (2) it should provide detailed descriptions with accurate localization of objects within the region. To address these two criteria, we propose a semantic score and localization score to rank the responses. The descriptions with the highest and the lowest scores are paired to form preference data for DPO training.

Semantic Score. We introduce the semantic score to evaluate the relevance between the generated descriptions and the semantics of queried image regions. We leverage a pre-trained CLIP model [42] to compute the cosine similarities of text and visual embeddings as defined in Eq. (1):

$$S(I,T) = \alpha * \cos(\mathcal{F}_{region}(I), \mathcal{F}_{text}(T)) \quad (1)$$

Where α is the scale of similarities, which is set as 5 in our work to balance the range of the semantic score, I and T are the input image and MLLM generated region description with grounding text removed; $\mathcal{F}_{region}$ and $\mathcal{F}_{text}$ denotes the visual embedding for image region and text embeddings, respectively.

When extracting image region embeddings, a straightforward approach is to crop the image region I_{crop} and directly extract visual embeddings. However, the similarity score with such embedding tends to overly focus on the region's details while neglecting the image's surrounding context. To address this limitation, we supplement it with similarities S_{local} from visual embeddings of intact images

that incorporate local attention. Specifically, we feed the original image into CLIP and replace the final layer of the vision encoder that aggregates the embeddings with a local-attention layer. This modification allows the model to better account for the context around the region of interest. As defined in Eq. (2), we then use the average of the cropped image's similarity score and the full image's similarity score with local attention as the final semantic score, which effectively evaluates the extent of fine-grained alignment between the region description and local visual semantics.

$$S_{sem} = \frac{1}{2}(S(I_{crop}, T) + S_{local}(I, T)) \tag{2}$$

Localization Score. We propose a localization score to evaluate how detailed the MLLM responds in describing objects within the queried region and its grounding accuracy. This score is calculated based on the number of objects mentioned in the description that match the ground truth objects in the region. In practice, we use Grounding DINO [34] and the cropped image region to extract bounding boxes for objects mentioned in the description. The extracted objects are then combined with the original object annotations to form a set of ground truth objects within the region. Next, we extract the grounding results from MLLM-generated descriptions and combine them with the results from Grounding DINO to form the predicted objects. Finally, we compute the average IoU between the predicted objects and the ground truth as the localization score. The detailed process is outlined in Algorithm 1.

The localization score encourages the model to include more detailed descriptions for involved objects and accurately localize them in its responses. Finally, we combine the semantic and localization scores for each grounded region description:

$$S = \lambda S_{sem} + (1 - \lambda) S_{loc} \tag{3}$$

where λ is set to 0.8 in our implementation. Then, the descriptions with the highest and lowest scores are paired as preferred and rejected data for preference optimization.

Grounded Region Description Refinement. After obtaining the preference data pairs, we further enhance the divergence of the grounding results of the preferred and rejected descriptions to encourage the model to distinguish between accurate and inaccurate object localization. To achieve this, we refine the grounding results in the preferred descriptions while keeping the rejected ones unchanged. In practice, we leverage the results obtained while computing the localization score, including the object box predictions B_{pred} and ground-truth object boxes B_{gt}. We retain only those predictions that match the ground truths (IoU > 0.5) and replace their bounding boxes with the matched ones. Then, we remove duplicates of predictions based on their textual position in the description and IoUs. Finally, we reinsert the refined object box predictions into the region description, resulting in an improved grounded region description with more precise coordinates.

Algorithm 1 Computing Localization Score

Input: Cropped Image Region I_{crop}, Grounded Region Description T generated by MLLMs, Object Bounding Box Annotations B_{anno} for the Queried Region.

Output: Localization Score: S_{loc}.

1: Extract bounding boxes B_{text} from the description T and get the plain text T_{plain}.

2: Leverage Grounding DINO to get grounded object results B_{ground} from T_{plain}.

3: Get the set of ground truth object boxes B_{gt} by aggregating B_{ground} and B_{anno} and removing duplicated boxes.

4: Get the set of object box predictions B_{pred} for the description T by aggregating B_{ground} and B_{text} and removing duplicated boxes.

5: Computing IoU matrix $\mathbf{m}[i, j] = IoU(B^i_{gt}, B^j_{pred})$ between B_{gt} and B_{pred}.

6: Filter the IoU by a threshold of 0.5.

$$\mathbf{p}[i, j] = \begin{cases} 0 & \mathbf{m}[i, j] < 0.5 \\ \mathbf{m}[i, j] & \text{otherwise} \end{cases}$$

7: **return** $S_{loc} = \frac{1}{n} \sum_i^n \max_j \mathbf{p}[i, j]$

3.3. Preference Optimization

After curating the preference dataset, we finetune MLLMs through DPO and adopt LORA to save the training cost. The loss for optimizing MLLMs is defined as:

$$\mathcal{L} = -\mathbb{E}_{(x, y_w, y_l)}\left[\log \sigma(\beta \log \frac{\pi_*(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_*(y_l|x)}{\pi_{\text{ref}}(y_l|x)})\right] \tag{4}$$

where y_w and y_l are the preferred and rejected description data; $\pi_ref(y|x)$ is the base reference policy model, i.e., the initial instruction-tuned MLLM which is frozen during the training; $\pi_*(y|x)$ denotes the policy model which inherits from the instruction-tuned model with its LORA weights updated in the training process.

4. Experiments

4.1. Experiment Setups.

Implementation Details. In this work, we experiment with the proposed SPR with three MLLMs with spatial understanding capabilities, including Ferret [58], LLava-OneVision [24], and CogVLM-Grounding [52]. To construct preference data, we randomly select 10k images with object annotations from the training set of Objects365

Table 1. Experiments on the Referring Expression Comprehension task (Acc@0.5) on datasets RefCOCO/+/g , and the Phrase Grounding task (Recall@1) on Flickr30k Entities dataset. "-" indicates results are unavailable or that MLLMs do not support multi-object grounding.

Method	RefCOCO			RefCOCO+			RefCOCOg		Flickr30k Entities	
	val	testA	testB	val	testA	testB	val	test	val	test
UNITER [12]	81.41	87.04	74.17	75.90	81.45	66.70	74.02	68.67	–	–
UniTAB [55]	86.32	88.84	80.61	78.70	83.22	69.48	79.96	79.97	78.76	79.58
MDETR [21]	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	82.3	83.8
MiniGPT-v2-7B [70]	88.06	91.29	84.30	79.58	85.52	73.32	84.19	84.31	-	-
VistaLLM [41]	88.1	91.5	83.0	82.9	89.8	74.8	83.6	84.4	-	-
LLaVA-Grounding [64]	89.16	-	-	81.68	-	-	84.82	-	83.03	83.62
Shikra-7B [10]	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19	75.84	76.54
Shikra-13B [10]	87.83	91.11	81.81	82.89	87.79	74.41	82.64	83.16	77.41	78.44
Griffon-13B [60]	89.4	92.5	84.6	83.3	88.4	76.0	85.1	86.1	83.7	84.2
LLava-OV-7B [24]	74.77	82.59	64.04	70.17	79.85	58.48	72.34	71.39	-	-
+ SPR	76.66	82.52	65.97	71.62	79.87	59.99	72.98	71.55	-	-
Ferret-7B [58]	87.49	91.35	82.45	80.78	87.38	73.14	83.93	84.76	80.39	82.21
+ SPR	88.39	91.67	83.91	82.07	87.84	74.19	85.58	85.75	81.53	83.34
Ferret-13B [58]	89.48	92.41	84.36	82.81	88.14	75.17	85.83	86.34	81.13	84.76
+ SPR	89.94	93.06	85.12	83.29	88.89	75.74	86.46	86.92	81.82	83.75
CogVLM-Grounding-17B [52]	92.76	94.75	88.99	88.68	92.91	83.39	89.75	90.79	-	-
+ SPR	92.95	94.87	89.15	88.83	92.95	83.84	90.01	90.96	-	-

Table 2. Experiments on Referring Expression Comprehension task under different IoU thresholds. The results are the average on RefCOCO, RefCOCO+, and RefCOCOg datasets.

IoU Threshold	0.5	0.6	0.7	0.8	0.9
Ferret-7B	83.91	81.28	76.72	67.02	43.25
+SPR	84.93	82.36	78.42	70.09	52.21
Ferret-13B	85.56	82.94	78.57	70.04	49.55
+SPR	86.18	83.63	79.93	72.03	53.61

Dataset [46], then construct random regions to query models to generate grounded region descriptions. We adopt LORA [17] for tuning MLLMs. The training is conducted on one A100 GPU, which takes around 3 and 5 hours for Ferret 7B and 13B models, respectively. Please refer to Appendix for more details on preference data construction and hyperparameter selection.

Evaluation Benchmarks We evaluate our method on three types of benchmarks: (1) Grounding tasks that evaluate the localization accuracy, including referring expression comprehension (REC) and phrase grounding; (2) Region description task on Refcocog [22] and visual genome [23], and Ferret Bench [58] for comprehensive spatial understanding; (3) General benchmarks TextVQA [47], GQA [19], LLaVA-Bench [33], and hallucination benchmark POPE [56].

4.2. Experiments on REC

We first evaluate our method on the referring expression comprehension (REC) task on RefCOCO [22], RefCOCO+ [22], and Refcocog [36]. The task requires the model to locate the object or region given a short description, which evaluates the model's fine-grained visual grounding abilities under the single-object referent scenarios. As shown in Tab. 1, our proposed SPR framework consistently improves the performance of three baseline MLLMs on different datasets for all model sizes. Considering that the REC results are based on an IoU threshold of 0.5, the improvement on its performance indicates that the model localized more objects successfully. Hence, this improvement can be largely attributed to the introduction of localization scores when constructing the preference data in SPR. Region descriptions that accurately mention more objects could achieve higher localization scores in SPR and be more likely to serve as preferred data, facilitating the model to attend to more objects and their locations in the image.

To better evaluate the impact of SPR on the localization capability of MLLMs, we also conduct REC experiments with higher IoU thresholds by gradually increasing the IoU threshold of valid REC results from the default value of 0.5 to 0.9. As shown in Tab. 2, the improvements brought by SPR significantly increase as the threshold rises, with accuracy gains of 8.96 and 4.06 for the 7B model and the 13B model, respectively, when the IoU threshold rises to 0.9. With SPR, the localization accuracy of the objects in the model's response is greatly improved. Equipped with the grounded region description refinement and the supervision of preferred-rejected localization data in SPR, the model can respond more accurately to grounding object locations, demonstrating the effectiveness of incorporating preference optimization for region description and object localization in the fine-grained spatial understanding of MLLMs.

Table 3. Experiments on Phrase Grounding task under different IoU thresholds. The results are averaged over the validation and test set of the Flickr30k dataset.

IoU Threshold	0.5	0.6	0.7	0.8	0.9
Ferret-7B	81.3	76.14	67.55	53.86	29.98
+SPR	82.44	77.14	69.25	56.19	33.99
Ferret-13B	82.94	76.62	68.34	55.74	32.60
+SPR	82.78	77.23	69.74	56.96	34.18

Table 4. Experiments on the region captioning task on Refcocog and Visual Genome datasets.

Method	Refcocog		Visual Genome	
	METEOR	ROUGE_L	METEOR	ROUGE_L
Ferret-7B	12.3	15.6	17.4	29.6
+SPR	13.5	20.4	17.6	29.7
Ferret-13B	12.9	26.4	17.9	31.0
+SPR	13.3	27.2	18.2	31.3

Table 5. Experiments on the Ferret Bench. "Description", "Reasoning", and "Grounding" denote the Referring Description, Referring Reasoning, and Grounding in Conversation tasks.

Model	Description	Reasoning	Grounding	Avg.
Ferret-7B	68.7	67.3	57.5	64.5
+ SPR	70.0	68.4	58.1	65.5
Ferret-13B	70.6	68.7	59.7	66.3
+ SPR	70.8	72.6	60.2	67.9

Table 6. Experiments on the general and hallucination benchmarks. We report the accuracy for GQA and VQA and the F1 score on POPE.

Model	VQA^T	GQA	LLaVA	POPE
Ferret-7B	-	-	64.7	85.36
+ SPR	-	-	66.3	85.69
LLaVA-OV-7B	75.89	62.21	88.9	88.12
+ SPR	76.07	62.42	91.4	88.49

4.3. Experiments on Phrase Grounding

Furthermore, we experiment with the phrase grounding task on Flickr30k Entity [40]. In phrase grounding, the queried object phrases are combined in a single question, requiring MLLMs to detect the locations of multiple objects in a single response, which makes it more challenging than the single-object referring task like REC. Following [58], we adopt the question "What are the locations of [phrases]?" and evaluate the result using the MERGE-BOXES mode [21]. Since LLaVA-OneVision and CogVLM do not support multi-object detection, we report only the results for Ferret. As shown in Tab. 1, SPR effectively improves Ferret's performance in multi-object referent scenarios, especially for the 7B model, whose performance is even comparable to that of the 13B model.

We then experiment with the phrase grounding task under higher IoU thresholds. We found that the multi-object referencing setting in phrase grounding is more challenging than the single-object referencing in REC. As the IoU threshold increases, the performance drops more rapidly, indicating a significant demand for MLLM to improve the capabilities of more accurate localization. Our approach can significantly alleviate this issue. As shown in Tab. 3, SPR improves progressively with higher IoU thresholds, reaching a maximum gain of 4.01 and 1.58 Recall@1 for the 7B and 13B models, respectively. This experiment demonstrates the superiority of SPR in pursuing detailed descriptions with high-precision object localization.

4.4. Experiments on Region Captioning

Beyond the grounding task, we also verify our proposed SPR in improving the text qualities of MLLMs' outputs on fine-grained spatial understanding. We conduct experiments on the RefCOCOg and Visual Genome benchmarks. We prompt MLLMs with the question "Describe the region [region] in the image." to generate region captions and then evaluate the response quality using METEOR [5] and ROUGE_L [28] metrics. Tab. 4 shows that SPR effectively improves the quality of MLLM-generated region captions. After tuning with SPR, MLLMs are able to effectively attend to the user-specified regions and generate captions that better reflect the details of the region content.

4.5. Experiments on Ferret Bench

Ferret benchmark, proposed by [58], aims to evaluate MLLMs' fine-grained multimodal conversational capabilities such as referring description, referring reasoning, and grounded conversation. We follow the pipeline in [58] to prompt MLLMs with questions and employ GPT to evaluate the responses. As shown in Tab. 5, the proposed SPR can facilitate MLLMs in achieving better conversational qualities for fine-grained multimodal understanding, especially for the referring reasoning task, with an accuracy gain of about 3.9 for the 13B model. Equipped with SPR, MLLM can focus on more detailed visual information and generate responses that align better with human preferences.

4.6. Experiments on General Benchmarks

We further evaluate SPR on three general benchmarks to validate the benefits of improving MLLMs' spatial capabilities. TextVQA and GQA require MLLMs to answer questions or perform reasoning based on specific text, objects, and image content. LLaVA bench evaluates MLLMs comprehensive capabilities in conversation, description, and reasoning. As shown in Tab. 6, improving MLLMs' spa-

Table 7. Ablation Studies on the refinement of grounded region descriptions, and ratio λ between semantic and localization scores in ranking MLLM responses. We report the average results on Referring expression comprehension and phrase grounding tasks.

Method	REC	Phrase Grounding
Ferret-7B	83.91	81.30
+ SPR	84.93	82.44
w/o Refinement	84.41	91.38
$\lambda = 0.0$	84.25	81.83
$\lambda = 0.4$	84.34	81.95
$\lambda = 0.6$	84.66	82.13
$\lambda =$ **0.8**	84.93	82.44
$\lambda = 1.0$	84.45	81.87

Table 8. Ablation studies on the training strategy.

Method	REC	Phrase Grounding
Ferret-7B	83.91	81.30
+ Instruction Finetuning	84.35	81.72
+ DPO training	84.93	82.44

tial understanding capabilities consistently enhances their comprehension and reasoning abilities across diverse general scenarios, leading to performance gains on all three benchmarks. We also experiment on hallucination benchmark POPE, where SPR improves both Ferret and LLaVA-OneVision. This can be attributed to the preference data construction in SPR, where semantic and localization scores are applied to select region descriptions that better align with region content and reject those related to content outside the region or that contain hallucinations, thus effectively helping mitigate the hallucinations in MLLMs.

4.7. Ablation Studies

We conduct ablation studies over the two designs in SPR and evaluate the performance of Ferret-7B on the REC (Refcoco/+/g) and the Flickr30k phrase grounding tasks.

Score Ratio λ. In Sec. 3.2, we combine the semantic and localization scores to rank MLLMs generated descriptions with a score ratio λ. As shown in Tab. 7, we vary the λ from 0 to 1, and the trained models outperform the baseline model consistently. When λ equals zero, SPR achieves minimal gain, as the model might overly reward descriptions that simply list object names or fail to align with the region's semantics. On the other hand, when λ is set to 1, SPR completely disregards measuring how detail the MLLM describes the objects in the region. Under such situations, the model encourages coarse region descriptions with fewer objects involved and reduces the corresponding object bounding box texts in the preferred data, thereby hindering the training of MLLMs' localization capability. As the experiments show, a relatively high value of 0.8 achieves the best results and is set as the default value in SPR.

Refinement of Grounded Region Description. After constructing the preferred and rejected data pairs, we further refine the localization results in the preferred descriptions by completing bounding boxes for objects in the description that were not grounded and refining the existing bounding boxes. Tab. 7 shows the results of this refinement. We found that the refinement leads to greater improvements in the multi-object referring task of phrase grounding. This could be attributed to the fact that the baseline model often fails to follow instructions for providing bounding boxes for each mentioned object when generating region descriptions. After tuning by the refined descriptions, MLLMs could faithfully ground the mentioned objects, thereby improving the multi-object phrase grounding clearly.

4.8. Comparison with SFT

In this paper, we adopt DPO with accept-reject preference data to optimize MLLMs for spatial understanding, whereas prior work [10, 44, 61], primarily focuses on the stage of supervised instruction fine-tuning (SFT). In Tab. 8, we compare these two training approaches, where SFT is trained using only the accepted data. The results show that while SFT could improve MLLMs' localization capabilities, its performance gains are significantly lower than DPO. DPO optimizes MLLM by contrasting accepted and rejected data pairs, similar to the positive-negative sample training mechanism in traditional object detection algorithms. This approach helps models distinguish between accurate and inaccurate localizations and facilitates MLLM in spatial understanding more effectively. However, it is important to note that DPO training also relies on a well-trained SFT model as a foundation, making these two approaches complementary. In future work, we will further explore how to integrate SFT and DPO to enhance MLLMs' spatial understanding.

5. Conclusion

In this work, we propose SPR, a Spatial Preference Rewarding framework to enhance MLLM's fine-grained spatial understanding capabilities. We introduce a complete pipeline that includes (1) Constructing random region queries; (2) Prompting MLLMs to generate diverse grounded region descriptions; (3) Proposing semantic scores and localization scores to rank the descriptions comprehensively; (4) Refining the localization quality of preference data; (5) Finetuning MLLMs to optimize against detailed and accurate spatial understanding. The entire framework does not require additional human labor or external MLLMs, with minimal overhead on training costs. SPR addresses the lack of direct optimization for positive and negative localization samples in MLLM training, enhancing their localization capabilities and promoting better alignment with human preferences. Experiments demonstrate that SPR significantly improves MLLMs' performance on standard referring and grounding tasks for spatial understanding.

Acknowledgement

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).
This study is also supported by the MOE Tier-2 project, with project number MOE-T2EP20123-0003.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 3

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

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1

[4] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 3

[5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 7

[6] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. 3

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with Transformers. In *ECCV*, 2020. 1, 3

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 1

[9] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 3

[10] 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, 3, 6, 8

[11] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 3

[12] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 6

[13] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024. 3

[14] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 1, 3

[15] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[16] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13806, 2024. 1, 3

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 6

[18] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. 1

[19] 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

[20] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 3

[21] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021. 6, 7

[22] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 3, 6

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 3, 6

[24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 5, 6

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 3

[26] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023. 3

[27] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6657–6678, 2024. 3

[28] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 7

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 1

[30] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1

[32] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 3

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3, 6

[34] 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. 5

[35] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer, 2025. 1, 3

[36] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 6

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 3

[38] Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. *arXiv preprint arXiv:2408.10433*, 2024. 1, 3

[39] 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. 3

[40] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 7

[41] Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14076–14088, 2024. 6

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3, 4

[43] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[44] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12977–12987, 2024. 3, 8

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[46] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 6

[47] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Con-*

ference on Computer Vision and Pattern Recognition, pages 8317–8326, 2019. 6

[48] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 1, 3

[49] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 3

[50] Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*, 2024. 1, 3

[51] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 3

[52] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1, 5, 6

[53] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3

[54] Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. $beta$-dpo: Direct preference optimization with dynamic $beta$. *arXiv preprint arXiv:2407.08639*, 2024. 4

[55] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022. 6

[56] Li Yifan, Du Yifan, Zhou Kun, Wang Jinpeng, Zhao Wayne Xin, and Wen Ji-Rong. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 3, 6

[57] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, and Jing Shao. Context and attribute grounded dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6241–6250, 2019. 3

[58] 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, 3, 5, 6, 7

[59] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arxiv*, 2023. 1, 3

[60] Yufei Zhan, Yousong Zhu, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon v2: Advancing multimodal perception with high-resolution scaling and visual-language co-referring. *arXiv preprint arXiv:2403.09333*, 2024. 3, 6

[61] Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. In *European Conference on Computer Vision*, pages 405–422. Springer, 2025. 1, 3, 8

[62] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024. 1

[63] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Chunyuan Li, Jainwei Yang, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2025. 3

[64] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Chunyuan Li, Jainwei Yang, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2025. 6

[65] Mengxi Zhang and Kang Rong. Automated multi-level preference for mllms. *arXiv preprint arXiv:2405.11165*, 2024. 1, 3

[66] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 3

[67] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6578–6587, 2019. 3

[68] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024. 3

[69] Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*, 2024. 1, 3

[70] 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, 3, 6