Advancing Fine-Grained Visual Understanding with Multi-Scale Alignment in Multi-Modal Models

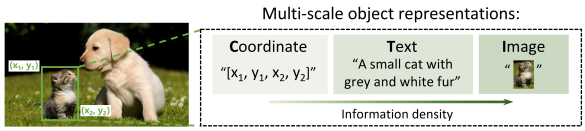
Wei Wang[*1,2], **Zhaowei Li**[*2,3], **Qi Xu**[2], **Linfeng Li**[2], **Yiqing Cai**[2], **Botian Jiang**[2,3], **Hang Song**[2], **Xingcan Hu**[1], **Pengyu Wang**[3], **Li Xiao**[1]
[1]MoE Key Laboratory of Brain-Inspired Intelligence Perception and Cognition
University of Science and Technology of China, [2]ByteDance Inc, [3]Fudan University
wangweiii@mail.ustc.edu.cn, lizhaowei126@gmail.com

Abstract

Multi-modal large language models (MLLMs) have achieved remarkable success in fine-grained visual understanding across a range of tasks. However, they often encounter significant challenges due to inadequate alignment for fine-grained knowledge, which restricts their ability to accurately capture local details and attain a comprehensive global perception. While recent advancements have focused on aligning object expressions with grounding information, they typically lack explicit integration of object images, which contain affluent information beyond mere texts or coordinates. To bridge this gap, we introduce a novel fine-grained visual knowledge alignment method that effectively aligns and integrates multi-scale knowledge of objects, including texts, coordinates, and images. This innovative method is underpinned by our multi-scale fine-grained enhancement data synthesis pipeline, which provides over 300K essential training data to enhance alignment and improve overall performance. Furthermore, we present TinyGroundingGPT, a series of compact models optimized for high-level alignments. With a scale of approximately 3B parameters, TinyGroundingGPT achieves outstanding results in grounding tasks while delivering performance comparable to larger MLLMs in complex visual scenarios. The data and code will be released in https://github.com/wwangweii/TinyGroundingGPT.git.

1 Introduction

Recent advancements in multi-modal large language models (MLLMs) have showcased remarkable capabilities in multi-modal understanding, reasoning, and interaction, garnering unprecedented attention (Touvron et al., 2023; Bai et al., 2023; Li et al., 2025b,a; Zheng et al., 2025b,a; Jian et al., 2025a, 2024). MLLM research in fine-grained visual understanding has advanced significantly, particularly through early contributions from Shikra (Chen et al., 2023a) and Kosmos-2 (Peng et al., 2023) in textually formatting positional vocabularies or object coordinates. Subsequent studies aimed at improving model performance primarily focused on common strategies, including parameter enlargement (Chen et al., 2023a; Peng et al., 2023; Li et al., 2024c; Bai et al., 2023) and dataset enrichment (Chen et al., 2023b; Bai et al., 2023; Wang et al., 2023; Chen et al., 2024b). Additionally, there is a growing interest in developing efficient, smaller fine-grained MLLMs (Li et al., 2024a; Hu et al., 2024; Yao et al., 2024; Zhu et al., 2023; Zhou et al., 2024) for real-world applications. Regardless of methods used, the core of fine-grained models lies in achieving better alignment between object texts and visual features, encompassing both coordinate and semantic information.



Models	C-T	C-I	T-C	T-I	I-T	I-C
SAM	✗	✗	✓	✗	✗	✓
LLaVA	✗	✗	✗	✗	✓	✗
Shikra, GroundingGPT	✓	✗	✓	✗	✓	✗
Qwen2-VL, InternVL2	✓	✗	✓	✓	✓	✗
Ours	✓	✓	✓	✓	✓	✓

Figure 1: The comparison of alignment for multi-scale object representations. The C, T, I denote object coordinates, texts and images respectively. The "X-Y" denote MLLMs handle input "X" and output "Y".

While effective, these methods face a significant challenge, i.e., the lack of fine-grained alignments. Visual objects typically encompass multi-scale representations with varying levels of information, including coordinates, texts, and images, as illustrated in Fig. 1. In this context, coordinates provide low-level object grounding information, texts offer primary descriptions that may not capture every detail, and images convey high-level information

*Equal contribution. Order is random.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 14271–14290
November 4-9, 2025

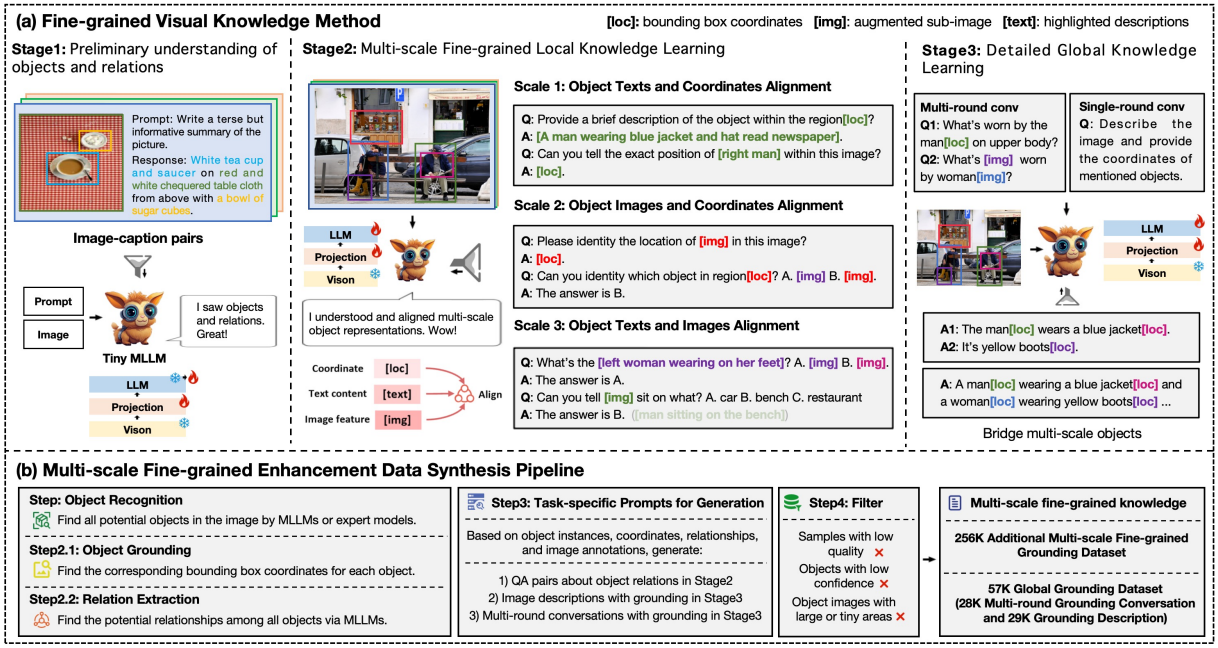


Figure 2: Illustration of the proposed multi-modal fine-grained visual knowledge alignment method. It adopts a three-stage training strategy that progresses from easy to hard and the multi-scale fine-grained enhancement data synthesis pipeline constructs over 300K fine-grained alignment data.

that extends beyond words. Most fine-grained models (Chen et al., 2023a; You et al., 2023; Li et al., 2024c) primarily focus on alignments between object texts and coordinates (i.e., T-C and C-T), often neglecting direct interactions with object images. Although recent models like Qwen2-VL (Wang et al., 2024a) and InternVL2 (Chen et al., 2024b) can process multiple image inputs and understand relationships between the main image and object images (T-I), they still struggle to establish explicit alignments among object texts, coordinates, and images. This limitation can lead to hallucinations and insufficient grounding capabilities (Chen et al., 2024a).

To achieve high-level alignments and integrate multi-granularity knowledge, as illustrated in Fig. 2(a), we introduce a fine-grained visual knowledge alignment method that effectively aligns object texts, coordinates, and images across multiple scales. Our method adopts a three-stage training strategy that progresses from easy to hard: 1) Object and Relation Perception Pretraining: To develop a foundational understanding of object texts and images, we implement a progressive training approach for MLLMs based on a pretrained LLM. 2) Multi-scale Fine-grained Local Knowledge Alignment: To attain fine-grained visual understanding and share multi-scale object knowledge, we conduct data-driven high-level alignments among object text descriptions, bounding box coordinates, and image features. 3) Detailed Global Knowledge Alignment: To enhance the model's global understanding by integrating fine-grained knowledge, we guide the MLLMs to bridge different objects with multi-scale representations. To support this method, we propose a multi-scale fine-grained enhancement data synthesis pipeline (see Fig. 2(b)) that constructs alignment data from both local and global perspectives. Leveraging this framework, we propose TinyGroundingGPT, which requires less storage for deployment while outperforming larger models across multiple benchmarks, particularly in grounding tasks. Our contributions can be summarized as follows:

- We introduce a fine-grained visual knowledge alignment method that enables the model to progressively enhance its fine-grained visual understanding through both global and local multi-scale object alignments.
- We develop a multi-scale fine-grained enhancement data synthesis pipeline that leverages open-source datasets and advanced models to generate over 300K essential training data for fine-grained alignment.
- We introduce TinyGroundingGPT, a series of compact models with 1.5B and 3B parameters, which excel in multi-modal understanding and grounding capabilities, achieving per-

formance comparable to larger 7B MLLMs.

2 Related Works

Multi-modal Large Language Models Recent progress in large language models (LLMs) such as ChatGPT and LLaMA (Touvron et al., 2023) has spurred the development of multi-modal LLMs. Notable models like GPT-4V (OpenAI, 2023) have demonstrated strong multi-modal capabilities in visual tasks. Early open-source models, including BLIP-2 (Li et al., 2023a), MiniGPT-4 (Zhu et al., 2023), and LLaVA (Liu et al., 2024), leverage pre-trained LLMs and perform well in visual question answering. Subsequent models, such as Qwen-VL (Bai et al., 2023), InternVL (Chen et al., 2024b), and MiniCPM-V (Yao et al., 2024), further enhance capabilities through dynamic resolution, expanded training data, and reinforcement learning, achieving notable results in OCR and grounding while improving response credibility.

However, the high parameter counts of MLLMs lead to significant training and deployment costs, limiting their widespread use. To address this, lightweight LLMs such as Mini-Gemini (Li et al., 2024a), MobileVLM (Chu et al., 2024), and MiniCPM-V (Yao et al., 2024) have been developed. These models, combined with optimized structures and training strategies, achieve performance comparable to larger models. Additionally, studies (Hsieh et al., 2023; Wang et al., 2024b; Shu et al., 2024) have explored distilling capabilities from larger models to enable smaller models to acquire complex reasoning abilities.

Fine-grained Multi-modal Models Recent works have focused on MLLMs for fine-grained understanding, with applications in tasks like grounding and OCR. Region-level understanding MLLMs (Yuan et al., 2024; Guo et al., 2024; Lu et al., 2023; Jian et al., 2025b) achieved local alignment between specific region features and texts. Methods such as Shikra (Chen et al., 2023a) and Kosmos-2 (Peng et al., 2023) enhanced visual grounding by constructing datasets with coordinate information, often converting visual tasks into instruction-following formats using templates. Other approaches integrated additional visual components, like GLaMM (Rasheed et al., 2024) and LLaVA-Grounding (Zhang et al., 2023b), or extracted regional features as supplementary inputs, as seen in Ferret (You et al., 2023), NExT-Chat (Zhang et al., 2023a), and GPT4RoI (Zhang et al., 2023c). GroundingGPT (Li et al., 2024c) extended support for multi-modal grounding tasks. Models like VisionLLMv2 (Wu et al., 2024) and UnifiedMLLM (Li et al., 2024b) expanded capabilities for various visual tasks, including image editing and segmentation. For fine-grained tasks, models such as LLaVA-UHD (Xu et al., 2024) and Qwen2-VL (Wang et al., 2024a) explored dynamic high-resolution techniques, improving OCR results. However, these models often lack systematic alignment among object texts, coordinates, and images, limiting the integration of multi-scale representations.

3 Method

In this paper, we first introduce a novel fine-grained visual knowledge alignment method that harnesses the potential of MLLMs by aligning object texts, coordinates, and images across multiple scales, as shown in Fig. 2(a). Our method consists of three training stages that progress from easy to hard: (a) Object and Relation Perception Pretraining, which enables the model to understand multimodal inputs, identifying objects in images and their interrelations. (b) Multi-scale Fine-grained Local Knowledge Alignment by which the model is guided to achieve multi-scale, fine-grained alignments, accommodating diverse inputs such as object texts, coordinates, and images. (c) Detailed Global Knowledge Alignment which focuses on model training for global alignment and understanding, further integrating fine-grained information and bridging different objects with multi-scale representations. To support this high-level alignment, we then propose a multi-scale fine-grained enhancement data synthesis pipeline, as illustrated in Fig 2(b), which generates multi-scale alignment datasets from both global and local perspectives. Building on this framework, we propose Tiny-GroundingGPT, which requires less storage for deployment while outperforming larger parameter models across multiple benchmarks, particularly in hallucination evaluation and grounding tasks.

3.1 Fine-grained Visual Knwoledge Alignment

We elaborate the three training stages in our fine-grained visual knowledge alignment method below.

Object and Relation Perception Pretraining In this stage, we aim for the model to comprehend

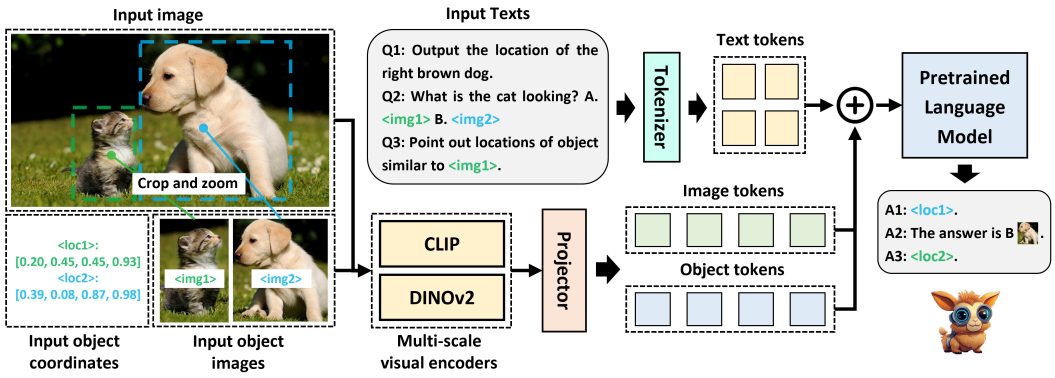


Figure 3: The model architecture of our proposed TinyGroundingGPT. It utilizes multi-scale visual encoders and supports queries regarding different object representations. Object images are cropped and zoomed from the input image according to the input coordinates.

multi-modal inputs, recognizing the objects present in the image and the relationships among them, which forms the foundation for subsequent reasoning and grounding tasks. Throughout the training process, we initially keep the LLM and encoder frozen, training only the projector to connect the text and image semantic spaces. Subsequently, we train both the LLM and the projector to enhance the understanding of objects and their relationships. We utilize LLaVA-Pretrain-595k (Liu et al., 2024) and each sample is accompanied by a sampled instruction that requires the model to provide a concise description of the image.

Multi-scale Fine-grained Local Knowledge Alignment After the initial training stage, where the model learns to recognize objects and their relationships, it still lacks the grounding capability to accurately locate these objects in images and to integrate different representations of a single object. In this stage, we therefore train the model to achieve fine-grained alignments among object texts, coordinates, and images, fully sharing their multi-scale knowledge for each representation. We utilize original visual grounding datasets such as RefCOCO (Kazemzadeh et al., 2014), RefCOCO+ (Kazemzadeh et al., 2014), RefCOCOg (Mao et al., 2016) and Visual Genomes (Krishna et al., 2017), along with a developed multi-scale fine-grained enhancement data synthesis pipeline (details provided in the following subsection) to construct a fine-grained grounding dataset. The instances in the training data can be categorized into three classes:

- Object **Texts** and **Coordinates** Alignment: The model refers to corresponding coordinates for a given object text description or describes a region based on input coordinates.
- Object **Images** and **Coordinates** Alignment: Given an augmented object image, the model identifies its location within the image. When provided with coordinates, the model selects the most relevant object images.
- Object **Texts** and **Images** Alignment: The model selects the most relevant augmented object image based on the input question or answers inquiries about the relationships involving augmented object images.

Throughout the training process, we train both the LLM and the projector. Afterwards, the model can effectively perform fine-grained image understanding by achieving high-level alignments among object texts, images, and coordinates, while sharing multi-scale knowledge across each representation.

Detailed Global Knowledge Alignment Despite achieving a fine-grained understanding of multi-modal data in the previous stage, the model lacks systematic training for global image comprehension and the ability to connect different objects with varied representations. Specifically, in the previous stage, only the representations of individual objects in each training sample were aligned. In this stage, our goal is to further align and integrate multiple objects within a single image input to enhance global knowledge learning. To achieve this, in addition to utilizing common image annotation datasets for instruction tuning, including LLaVA-v1.5-mix665k (Liu et al., 2024) and ShareGPT4V (Chen et al., 2023b), we construct a global grounding dataset with high-level

fine-grained alignments based on Flickr30K Entities (Plummer et al., 2015): 1) Multi-round Grounding Conversation Data: This dataset guides the model to achieve a global understanding of the image through multi-round conversations, requiring it to combine fine-grained knowledge and thoroughly explore the relationships among different representations of various objects. 2) Grounding Description Data: This dataset prompts the model to provide a detailed description of the image to connect multi objects in one-round conversations, where the generated object texts are enhanced with coordinates to confirm their existence and effectively integrate grounding information.

This method enables us to leverage the fine-grained grounding alignment learned in the second stage to enhance the model's global grounding alignment. Additionally, we train both the LLM and the projector in this stage.

3.2 Multi-scale Fine-grained Enhancement Data Synthesis Pipeline

As shown in Fig. 2(b), we develop a multi-scale fine-grained enhancement data synthesis pipeline, and construct a multi-scale fine-grained grounding dataset (in Stage2) as well as a global grounding dataset (in Stage3). Specifically, given an image, we perform the following steps:

Object Recognition We employ expert models or MLLMs for object detection in input images, generating a list of identified objects, referred to as L_1. A prompt example for GPT-4V (OpenAI, 2023) is provided in Appendix Fig.7.

Object Grounding Beyond object text and coordinate pairs in original datasets such as RefCOCO, we apply grounding models to obtain bounding box coordinates. In this paper, we employ GroundingDINO (Liu et al., 2023b) to locate objects in L_1 and filter out those with low confidence, resulting in an object bounding box dictionary S_1.

Relationship Extraction To uncover the relationships between objects for subsequent QA generation, we instruct GPT-4V to identify potential connections among objects. As shown in Appendix Fig. 7, given the object list L_1, GPT-4V generates a list L_2 containing triples in the format $(\text{object1}, \text{relation}, \text{object2})$.

QA Generation Based on above L_1, S_1 and L_2, we use task-specific prompts for GPT-4V to generate different kinds of datasets (we provide case examples in Appendix Figs. 8 and 9): (1) 256K Additional Multi-scale Fine-grained Grounding Dataset: Compared to previous works (Li et al., 2024c; Chen et al., 2023a) that focused solely on the alignment between object texts and coordinates, we enhance the alignment format by constructing an additional multi-scale, fine-grained dataset. This dataset incorporates object images, texts, and coordinates, facilitating more fine-grained image understanding and multi-scale alignment. Details can be seen in Appendix Fig. 10. (2) 57K Global Grounding Dataset: To enhance the global alignment and bridge objects with various representations, we construct two kinds of datasets: 1) 28K Multi-round Grounding Conversation: This dataset includes multi-turn dialogue formats, focusing on point-to-point questions about local details. 2) 29K Grounding Description: This dataset features single-turn dialogue formats, emphasizing an understanding of overall image descriptions with fine-grained grounding information. See details in Appendix Figs. 11 and 12.

Filter We filter out QAs that contain object images with areas that are either too large or too small, as well as those with high Intersection over Union among object images in the options. Additionally, we exclude QAs related to objects with low confidence or those with an excessive number of bounding boxes. This helps avoid low-resolution noise or image reference ambiguity. Finally, we exclude low-quality QAs with the assistance of GPT-4V.

3.3 TinyGroundingGPT

Using our proposed alignment method and synthesis data, we train TinyGroundingGPT to demonstrate the effectiveness of our method. Fig. 3 illustrates the overall architecture of TinyGroundingGPT. Images in various formats are processed through multi-scale vision encoders to extract features. Specifically, we use the pre-trained visual encoders ViT-L/14 (Radford et al., 2021) and DINOv2-L/14 (Oquab et al., 2023) to extract image features, concatenating them to combine the global perception of CLIP and the local fine-grained understanding of DINOv2 (Jiang et al., 2023). These features are then mapped to the LLM embedding space using an MLP. In TinyGroundingGPT, the input supports both global images and object images, each represented by different special tokens: $< image >$ and $< object >$. These object images are cropped and zoomed from the global image based on the corresponding coordinates. We also support the input and output of object bounding box coordinates $< loc >$, represented in the text format $[x1, y1, x2, y2]$ with values in $[0.000, 1.000]$.

Type	Model	LLM Size	RefCOCO			RefCOCO+			RefCOCOg		Avg
			val	testA	testB	val	testA	testB	val	test	
Specialist	UNITER (Chen et al., 2020)	-	81.41	87.04	74.17	75.90	81.45	66.70	74.02	68.67	76.17
	MDETR (Kamath et al., 2021)	-	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	81.81
	UniTAB (Yang et al., 2022)	-	86.32	88.84	80.61	78.70	83.22	69.48	79.96	79.97	80.89
Generalist	KOSMOS-2 (Peng et al., 2023)	1.6B	52.32	57.42	47.26	45.48	50.73	42.24	60.57	61.65	52.21
	Shikra (Chen et al., 2023a)	7B	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19	82.93
	NExT-Chat* (Zhang et al., 2023a)	7B	85.50	90.00	77.90	77.20	84.50	68.00	80.10	79.80	80.38
	Ferret* (You et al., 2023)	7B	87.49	91.35	82.45	80.78	**87.38**	73.14	**83.93**	**84.76**	83.91
	GroundingGPT (Li et al., 2024c)	7B	88.02	91.55	82.47	81.61	87.18	73.18	81.67	81.99	83.46
	InternVL2$^+$ (Chen et al., 2024b)	2B	82.3	88.2	75.9	73.5	82.8	63.3	77.6	78.3	77.74
	Qwen2-VL$^+$ (Wang et al., 2024a)	2B	87.6	90.6	82.3	79.0	84.9	71.0	81.2	80.3	82.11
Generalist	TinyGroundingGPT	3B	**89.16**	**92.24**	**85.38**	**81.70**	87.16	**75.09**	83.27	84.08	**84.76**
		1.5B	86.76	90.42	81.81	78.86	84.65	70.24	79.88	80.04	81.58

Table 1: Performance comparison on the referring expression comprehension(REC) task. "*" indicates that the model employs additional image region perception modules and "+" indicates that the model uses dynamic high-resolution. The best results are highlighted in bold, while the second-best results are underlined.

4 Experiments

4.1 Experimental Setup

We employ Qwen2.5-3B and Qwen2.5-1.5B (Yang et al., 2024) as the language models for our TinyGroundingGPT. During the training process, all images were padded to a square shape and resized to a resolution of 336 × 336. For more details on hyper-parameter settings, training processes and datasets, please refer to the Appendix A.1 and A.2.

4.2 Image Grounding Evaluation

To evaluate the image grounding capability of TinyGroundingGPT, we conducted experiments on the Reference Expression Understanding (REC) task, which involves locating the bounding box for a given text reference. We utilized three datasets: RefCOCO, RefCOCO+, and RefCOCOg. We compared TinyGroundingGPT against various baseline models, including end-to-end multi-modal models such as UNITER (Chen et al., 2020), MDETR (Kamath et al., 2021), and UniTAB (Yang et al., 2022), as well as LLM-based models like KOSMOS-2, Shikra, NExTChat, Ferret, and GroundingGPT. Additionally, smaller models such as InternVL2 and Qwen2-VL were included. We used a unified prompt formatted as "Output the coordinate of $< \text{exp} >$", where "$< \text{exp} >$" represents the reference expression. As shown in Table 1, TinyGroundingGPT demonstrates strong performance across all datasets, even with smaller LLM sizes (3B and 1.5B), matching or exceeding the performance of specialized fine-tuned models and larger MLLMs with additional image perception modules. Notably, the 3B model achieved state-of-the-art results on several benchmarks, attaining the highest average accuracy. Furthermore, TinyGroundingGPT-1.5B showed comparable grounding results, outperforming Next-Chat-7B on nearly all test sets.

4.3 Image Understanding Evaluation

We evaluated TinyGroundingGPT on seven benchmarks, providing a comprehensive assessment of its performance across various metrics. As shown in Table 2, TinyGroundingGPT-3B achieves results comparable to models such as MiniCPM-V-2 and InternVL-2, which utilize dynamic high resolution or enriched training data. Compared to models with similar fine-tuning data, including LLaVA-1.5, GroundingGPT, TinyLLaVA, and LLaVA-Phi, TinyGroundingGPT-3B demonstrates superior image understanding capabilities on the VQA^{v2}, GQA, SQA, and POPE benchmarks, achieving increases of 2.6% on MMB and 1.2% on GQA over GroundingGPT-7B. Notably, TinyGroundingGPT-1.5B outperforms LLaVA-Phi, despite its larger parameter count, on most benchmarks. We further evaluated the MLLMs for object hallucination on the POPE benchmark, with more details provided in Appendix A.3. Overall, TinyGroundingGPT, optimized by our multi-scale visual knowledge alignment method, achieved impressive results across multiple evaluation sets.

4.4 Ablation Study

Ablation Study on Additional Multi-scale Fine-grained Grounding Dataset. In Stage 2, compared to traditional methods that rely solely on alignment datasets for object texts and coordinates, we utilized our constructed multi-scale

Models	LLM Size	VQA^{v2}	GQA	SQA^{I}	POPE	MME^{P}	MMB	LLaVA^{W}
BLIP-2 (Li et al., 2023a)	13B	41.0	41	61	85.3	1293.8	-	38.1
InstructBLIP (Dai et al., 2023)	7B	-	49.2	60.5	-	-	36	60.9
InstructBLIP (Dai et al., 2023)	13B	-	49.5	63.1	78.9	1212.8	-	58.2
Shikra (Chen et al., 2023a)	13B	77.4	-	-	-	-	58.8	-
LLaVA-1.5 (Liu et al., 2023a)	7B	78.5	62.0	66.8	85.9	**1510.7**	64.3	63.4
GroundingGPT (Li et al., 2024c)	7B	78.7	62.1	-	87.4	1454.2	63.8	**70.9**
Qwen-VL-Chat (Bai et al., 2023)	7B	78.2	-	68.2	-	1487.5	60.6	-
MiniCPM-V-2^{+} (Yao et al., 2024)	2.8B	-	-	-	87.8	-	**69.6**	69.2
InternVL-2^{+} (Chen et al., 2024b)	2B	-	61.0	-	**88.3**	1439.6	-	62.5
LLaVA-Phi (Zhu et al., 2024)	2.7B	71.4	-	68.4	86.7	1335.1	59.8	-
TinyLLaVA (Zhou et al., 2024)	2.7B	77.7	61.0	70.1	86.3	1437.3	68.3	67.1
TinyGroundingGPT	3B	**79.3**	**63.3**	**70.3**	87.9	1423.2	66.4	67.5
	1.5B	77.9	62.2	63.1	87.6	1392.4	64.2	65.3

Table 2: Comparison of MLLMs on image understanding benchmarks. Benchmark names are abbreviated due to space limits. VQA^{v2} (Goyal et al., 2017); GQA (Hudson and Manning, 2019); SQA^{I}:ScienceQA-IMG (Lu et al., 2022); POPE (Li et al., 2023b); MME (Fu et al., 2023); MMB:MMBench (Liu et al., 2025); LLaVA^{W}: LLaVA-Bench (In-the-Wild) (Liu et al., 2024). "+" indicates that the model uses dynamic high-resolution.

fine-grained grounding datasets for TinyGroundingGPT, enabling multi-scale alignment among object texts, images, and coordinates. The ablation study in Table 3 shows that our proposed multi-scale fine-grained alignment outperforms traditional referring data that only aligns object texts with coordinates. For both the 3B and 1.5B TinyGroundingGPT models, our method enhanced performance on the RefCOCO, RefCOCO+, and RefCOCOg benchmarks. For instance, on the RefCOCO+ benchmark, our method achieved an increase of 1.67% for the 3B model and 0.87% for the 1.5B model, demonstrating the effectiveness of our proposed fine-grained alignments and datasets.

Size	Alignment	RefCOCO	RefCOCO+	RefCOCOg
3B	T, C	87.35	78.89	83.25
	T, C, I	88.50	80.56	83.69
1.5B	T, C	85.93	77.05	79.54
	T, C, I	86.33	77.92	79.96

Table 3: Ablation study on our Additional Multi-scale Fine-grained Grounding Dataset in Stage 2. Here, T, C, and I denote Text, Coordinate, and Image, respectively. We report the average accuracy for each benchmark.

Ablation Study on Global Grounding Datasets. In Stage 3, we utilized the constructed Global Grounding Datasets for TinyGroundingGPT to bridge different objects with varied representations and enhance global image comprehension. The results presented in Table 4 showcased the effectiveness of this strategy. Notably, a reduction in hallucinations can be observed on the POPE benchmark. Overall, significant improvements in visual understanding benchmarks underscored the value of detailed global knowledge learning and the Global Grounding Datasets, which enhanced global object alignment by connecting different objects represented by texts, coordinates, and images.

Size	Global Align	GQA	VQA^{v2}	SQA	POPE	MMB
3B	✗	61.7	77.4	65.6	86.6	63.1
	✓	63.3	79.3	70.3	87.9	66.4
1.5B	✗	60.3	77.3	62.1	86.4	63.0
	✓	62.2	77.9	63.1	87.6	64.2

Table 4: Ablation study on our Global Grounding Datasets in Stage 3. If the model is trained without global alignment, it indicates that we do not use these datasets to further align different objects represented by texts, coordinates, and images.

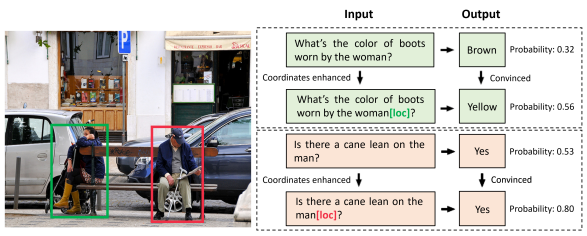


Figure 4: A case for the outputs of our TinyGroundingGPT when the input is either enhanced with coordinates or not. Probability values indicate the likelihood of generating corresponding tokens.

Ablation Study on Models. (1) Larger parameters: We applied our method to TinyGroundingGPT with the larger language model Qwen2.5-7B (Appendix A.4). (2) Vision encoders: We explored the effects of our multi-scale vision encoders (ViT and DINOv2) (Appendix A.5). (3) Qwen2 as pretrained

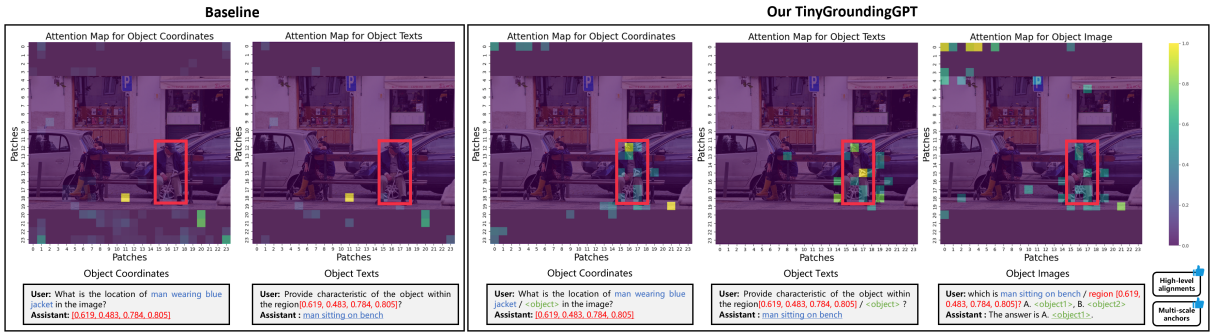


Figure 5: Visualization of the attention map for image patches with different object representation outputs (texts, coordinates, and images, underlined). The red bounding box denotes the target region. The attention at the four corners serves as anchors for grounding, while attention at specific objects highlights their importance.

LLM: We compared Qwen2.5 with Qwen2, demonstrating Qwen2.5's effectiveness (Appendix A.6). Overall, the results highlight the effectiveness of our proposed method and TinyGroundingGPT.

5 Discussion

5.1 Effectiveness of Fine-grained Knowledge

Our fine-grained visual knowledge alignment method not only improves grounding ability but also enhances comprehensive image understanding. Examples of image descriptions generated by TinyGroundingGPT are provided in Appendix Fig. 13, demonstrating the model's ability to avoid incorrect or nonexistent object descriptions. We evaluated the annotation quality by selecting 50 images from RefCOCO-test and using GPT-4V to score descriptions from different models. As detailed in Appendix A.7, TinyGroundingGPT-3B outperforms GroundingGPT-7B and Qwen2-VL-2B in overall quality and richness. Additionally, incorporating fine-grained knowledge into input questions for TinyGroundingGPT results in more accurate and persuasive responses. As shown in Fig. 4, adding coordinates to object texts in queries enhances response confidence compared to directly asking about objects in an image. This underscores the potential of fine-grained MLLMs.

5.2 Interpretability for High-level Alignments

Grounding MLLMs fundamentally model the maximum likelihood output based on visual inputs and text prompts. By conditioning on the referring prompt, the model identifies which parts of the image significantly influence the output. To demonstrate the effectiveness of our multi-scale fine-grained grounding capability, we visualize the attention map in the last layer of our TinyGroundingGPT. As shown in Fig. 5, the attention maps of our TinyGroundingGPT reveal distinct location attributions, unlike the baseline GroundingGPT-7B.

For object coordinates, high attention scores are concentrated at the four corners of the image, serving as anchors for bounding box coordinates, as well as at the locations of the intended objects mentioned in the prompt. When prompted to describe a specific region, the model directs increased attention to the corresponding object patches. For the output of an object image, the attention values between image patches and the target object highlight relevant regions and reinforce grounding anchors. This indicates that TinyGroundingGPT effectively learns both aligned features and grounding information for object images. In summary, our findings underscore the effectiveness of the proposed fine-grained visual knowledge alignment method, achieving high-level alignment among different object representations. This provides insights for further explaining MLLMs, particularly in grounding tasks. More visualizations can be found in Appendix Fig. 14.

6 Conclusion

In this paper, we introduce a novel fine-grained visual knowledge alignment method for MLLMs to address the limitations of fine-grained alignments in previous works. Our method progresses from easy to hard, emphasizing multi-scale fine-grained alignments among object texts, coordinates, and images from both local and global perspectives. This empowers models to effectively learn fine-grained knowledge and facilitates reasoning and grounding tasks. Additionally, we develop a multi-scale fine-grained enhancement data synthesis pipeline that leverages open-source datasets and advanced mod-

els to generate over 300K essential training samples. Building on this foundation, we train TinyGroundingGPT, a series of smaller models (1.5B and 3B parameters) optimized through high-level alignments, capable of handling various visual and grounding tasks, often surpassing larger models. Experimental results demonstrate the effectiveness of our proposed method and the generated datasets. Our work contributes to the advancement of practical applications for MLLMs.

Limitations

Our work has developed a fine-grained visual knowledge alignment method for MLLMs. Based on this, we constructed the necessary datasets and trained our proposed TinyGroundingGPT. There are several aspects that can be further improved: (1) Additional techniques can be applied to TinyGroundingGPT to further enhance its performance. For example, the dynamic high-resolution in works (Xu et al., 2024; Chen et al., 2024b) has been proved to improve image understanding capabilities. (2) Additional datasets, such as OCR datasets described in (Wang et al., 2024a), can be utilized for supervised fine-tuning of TinyGroundingGPT in Stage3 to further enhance its multi-modality capabilities. (3) Our current multi-scale alignments focus only on objects within a single image. This approach can be further extended to include similar objects or objects captured from different angles across multiple images, thereby enhancing the robustness and generality of alignment.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62202442.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

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

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023b. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Jinjie Gu, and Huajun Chen. 2024a. Unified hallucination detection for multimodal large language models. *arXiv preprint arXiv:2402.03190*.

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

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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*, pages 6904–6913.

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

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm:

Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709.

Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. 2025a. Look again, think slowly: Enhancing visual reflection in vision-language models. *Preprint*, arXiv:2509.12132.

Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. 2025b. Teaching vision-language models to ask: Resolving ambiguity in visual questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3619–3638.

Pu Jian, Donglei Yu, and Jiajun Zhang. 2024. Large language models know what is key visual entity: An llm-assisted multimodal retrieval for vqa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10939–10956.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. 2023. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*.

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

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. 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.

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

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

Junxian Li, Beining Xu, and Di Zhang. 2025a. Iag: Input-aware backdoor attack on vlms for visual grounding. *arXiv preprint arXiv:2508.09456*.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, et al. 2025b. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 415–423.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024a. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

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

Zhaowei Li, Wei Wang, YiQing Cai, Xu Qi, Pengyu Wang, Dong Zhang, Hang Song, Botian Jiang, Zhida Huang, and Tao Wang. 2024b. Unifiedmllm: Enabling unified representation for multi-modal multi-tasks with large language model. *arXiv preprint arXiv:2408.02503*.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. 2024c. 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.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

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

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer.

Junyu Lu, Dixiang Zhang, Songxin Zhang, Zejian Xie, Zhuoyang Song, Cong Lin, Jiaxing Zhang, Bingyi Jing, and Pingjian Zhang. 2023. Lyrics: Boosting fine-grained language-vision alignment and comprehension via semantic-aware visual objects. *arXiv preprint arXiv:2312.05278*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

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

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

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

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. 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.

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

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. 2024. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018.

Fangxun Shu, Yue Liao, Le Zhuo, Chenning Xu, Guanghao Zhang, Haonan Shi, Long Chen, Tao Zhong, Wanggui He, Siming Fu, et al. 2024. Llava-mod: Making llava tiny via moe knowledge distillation. *arXiv preprint arXiv:2408.15881*.

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

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

Wei Wang, Zhaowei Li, Qi Xu, Yiqing Cai, Hang Song, Qi Qi, Ran Zhou, Zhida Huang, Tao Wang, and Li Xiao. 2024b. Qcrd: Quality-guided contrastive rationale distillation for large language models. *arXiv preprint arXiv:2405.13014*.

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

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Wenhai Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, et al. 2024. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *arXiv preprint arXiv:2406.08394*.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. 2024. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

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

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

Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. 2023a. Next-chat: An

lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, et al. 2023b. Llava-grounding: Grounded visual chat with large multimodal models. *arXiv preprint arXiv:2312.02949*.

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

Xiaofan Zheng, Minnan Luo, and Xinghao Wang. 2025a. Unveiling fake news with adversarial arguments generated by multimodal large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7862–7869, Abu Dhabi, UAE. Association for Computational Linguistics.

Xiaofan Zheng, Zinan Zeng, Heng Wang, Yuyang Bai, Yuhan Liu, and Minnan Luo. 2025b. From predictions to analyses: Rationale-augmented fake news detection with large vision-language models. In *Proceedings of the ACM on Web Conference 2025*, pages 5364–5375.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.

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

Yichen Zhu, Minjie Zhu, Ning Liu, Zhiyuan Xu, and Yaxin Peng. 2024. Llava-phi: Efficient multi-modal assistant with small language model. In *Proceedings of the 1st International Workshop on Efficient Multimedia Computing under Limited*, pages 18–22.

A Appendix

A.1 Implementation Details

We present additional details about our experimental configuration to facilitate the reproduction of our model. The hyperparameters for all stages are summarized in Table 5. We adopted a progressive training strategy in Stage 1 because the loss after pretraining the MLP for TinyGroundingGPT was still relatively high (about 3.0). Further finetuning TinyGroundingGPT with both the MLP and pretrained LLMs reduced the loss to around 1.5 for improving multi-modality ability. Moreover, the object coordinates in training were normalized after padding.

Size	Stage 1		Stage 2	Stage 3
	Pretrain	Finetune		
Batch size	32	32	32	16
Learning rate	1e-3	2e-5	2e-5	2e-5
Epochs	1	1	1	2
Learning schedule	Cosine decay			
Warm-up ratio	0.03	0.03	0.03	0.03
Weight decay	0	0	0	0
BF16	✓	✓	✓	✓
TF32	✓	✓	✓	✓
DeepSpeed stage	ZeRO2			
GPUs	8xA100			

Table 5: The hyperparameters for model training.

A.2 Dataset Details

We provide additional details about the datasets we utilized, as summarized in Table 6. We also include additional examples of the generated datasets in Fig. 8 for Stage 2 and in Fig. 9 for Stage 3. Specifically, during the training process, the object images are cropped and zoomed from the original image and then fed into the vision encoders to obtain object image features. These features are subsequently used to replace the placeholder denoted as <img> in the QA pairs.

As described in Section 3.2, we developed a multi-scale fine-grained enhancement data synthesis pipeline, which includes the construction of a multi-scale fine-grained grounding dataset (in Stage 2) and a global grounding dataset (in Stage 3). In Fig. 7, we present the prompt messages used for object recognition and relation extraction to prepare additional data material. Fig. 10 illustrates the detailed processing steps involved in constructing the multi-scale fine-grained grounding dataset. Furthermore, Figs. 11 and 12 outline the processing steps for constructing the global grounding dataset. Additionally, Table 7 shows an example prompt used to evaluate the generated QA pairs. Based on these evaluations, we filtered out low-quality QAs (or descriptions), specifically those with incorrect answers or low-quality scores (<3).

A.3 Object Hallucination Evaluation

We evaluated MLLMs for object hallucination, as shown in Table 8. Higher accuracy and F1-score metrics, along with a lower 'Yes' metric, indicate better performance. Our TinyGroundingGPT yielded outstanding results across all three sampling subsets. Notably, TinyGroundingGPT-3B outperformed larger models like InstructBLIP-13B in the challenging Adversarial subset, achieving

Stage	Dataset		Samples
Stage1	LLaVA-Pretrain-595k		595K
Stage2	Alignment data	Text-coordinate pairs	4.1M
		Image-coordinate pairs	210K
		Text-image pairs	46K
Stage3	SFT data	LLaVA-v1.5, ShareGPT4V	665K
		Grounding-conv	28K
		Grounding-description	29K

Table 6: The dataset details used for model training.

You are tasked with evaluating QA pairs based on an image. Please assess the provided QA pairs according to the following criteria:
Quality (1-5):

1 - The QA is incoherent, lacks flow, and fails to convey the content of the image effectively.

2 - The QA is somewhat relevant but contains notable inaccuracies or lacks clarity.

3 - The QA is generally clear and relevant, though it may overlook some important details or context from the image.

4 - The QA is clear, coherent, and accurately reflects the content of the image, with only minor omissions.

5 - The QA is highly coherent and effectively captures the essence and details of the image, providing insightful and accurate information.

Please evaluate the given QA pair on a scale from 1 to 5 and provide a brief justification for your rating, as well as determine whether the QA is correct.

Your output should be structured as follows: "Quality": "Your rating here.", "Correct": "Yes or No", "Justification": "Your justification here.".

Do not include any additional text outside of this format.

Table 7: The prompt for GPT-4V to evaluate generated QAs.

an increase of 14.67% in accuracy and a 8.90% increase in F1 score, despite a decrease of 27.77% in the 'Yes' metric. Compared to GroundingGPT-7B, our 3B model excelled in the Popular and Adversarial subsets for both accuracy and F1 score. Similarly, TinyGroundingGPT-1.5B achieved higher accuracy and F1 score than some larger models like Shikra while maintaining a lower 'Yes' score. This superior performance can be attributed to its fine-grained knowledge alignment from both global and local perspectives during training.

A.4 Grounding Ability for Larger Model

We further apply our fine-grained visual knowledge alignment method to TinyGroundingGPT, using Qwen2.5-7B as the larger-parameter language model, to evaluate its image grounding capability. The results in Table 9 highlight the method's effectiveness, with increases of 0.68% on RefCOCO+-testA and 0.81% on RefCOCO+-testB.

A.5 Ablation Study for Vision Encoders

Image features from ViT-L/14 (Radford et al., 2021) (second-to-last layer) capture more object semantics, while those from DINOv2-L/14 (Oquab et al., 2023) (last layer) capture more local fine-grained details, as shown in Fig. 6. The multi-scale vision encoders in our proposed TinyGroundingGPT align well with the fine-grained alignment of our method. As shown in Table 10, this approach improves performance on benchmarks highly related to fine-grained understanding, such as POPE and VQA^{v2}.

A.6 Ablation Study for Pretrained LLM

We conduct additional experiments to explore the effect of the pretrained LLM on TinyGroundingGPT. As shown in Table 11, TinyGroundingGPT with Qwen2.5-1.5B outperforms that with Qwen2-1.5B in image understanding, highlighting the effectiveness of Qwen2.5 for our proposed TinyGroundingGPT.

A.7 Assessment for Image Annotation

As illustrated in Section 5.1, we provided examples of image descriptions generated by TinyGroundingGPT in Fig. 13. Moreover, we selected 50 images from RefCOCO-test and utilized GPT-4V to evaluate image descriptions produced by various methods. We assessed the image descriptions using scores ranging from 1 to 5 across three perspectives: "Quality," which reflects overall quality; "Richness," which measures the diversity of object descriptions; and "Accuracy," which pertains to precision. The prompt used for GPT-4V and the scoring details are presented in Table 12. As the results summarized in Table 13, TinyGroundingGPT achieved better overall quality and richness compared to GroundingGPT-7B and Qwen2-VL-2B.

Model	Alignment	RefCOCO+		
		val	testA	testB
TinyGroundingGPT-7B	T, C	83.98	88.08	77.90
	T, C, I	84.56	88.76	78.71

Table 9: Performance comparison on the referring expression comprehension(REC) task for whether conducting our proposed Multi-scale Fine-grained Local Knowledge Alignment.

Models	LLM Size	Random			Popular			Adversarial		
		Accuracy	F1-Score	Yes	Accuracy	F1-Score	Yes	Accuracy	F1-Score	Yes
LLaVA	7B	72.16	78.22	76.29	61.37	71.52	85.63	58.67	70.12	88.33
mPLUG-Owl	7B	53.97	68.39	95.63	50.90	66.94	98.57	50.67	66.82	98.67
MiniGPT-4	13B	79.67	80.17	52.53	69.73	73.02	62.20	65.17	70.42	67.77
InstructBLIP	13B	88.57	89.27	56.57	82.77	84.66	62.37	72.10	77.32	73.03
Shikra	7B	86.90	86.19	43.26	83.97	83.16	45.23	83.10	82.49	46.50
GroundingGPT	7B	89.79	89.22	43.13	88.23	87.38	43.23	86.17	85.50	45.43
TinyGroundingGPT	3B	**89.93**	**89.47**	43.08	88.56	**87.90**	43.43	**86.77**	**86.22**	45.26
	1.5B	89.59	88.98	42.92	**88.67**	**87.90**	42.87	86.74	86.04	44.77

Table 8: Results on the POPE benchmark for object hallucination evaluation. "Yes" represents the probability of positive answers to the given question.

Vision Encoder	VQA^{v2}	GQA	SQA^{I}	POPE	MME^{P}	MMB
ViT	76.5	61.4	64.8	85.8	1417.8	63.0
ViT + DINOv2	77.9	62.2	63.1	87.6	1392.4	64.2

Table 10: Ablation Study on our TinyGroundingGPT-1.5B for the multi-scale vision encoders.

Pretrained LLM	VQA^{v2}	GQA	SQA^{I}	POPE	MME^{P}	MMB
Qwen2-1.5B	76.3	61.2	56.8	85.7	1386.9	58.9
Qwen2.5-1.5B	77.9	62.2	63.1	87.6	1392.4	64.2

Table 11: Ablation Study on our TinyGroundingGPT-1.5B for the pretrained LLM.

A.8 More Visualizations

As illustrated in Section 5.2, we visualized the last-layer attention maps of both the GroundingGPT-7B baseline and our TinyGroundingGPT-1.5B. The attention map in grounding MLLMs not only enhances interpretability but also illustrates the alignment between the model's output and the input image. The process for obtaining the heatmap of attention involves several steps: (1) we select the attention scores between image patches and object representations (i.e., texts, coordinates, and images); (2) we sum the attention scores across the dimensions of both the attention heads and object representations; (3) We map the normalized attention scores onto the input image patches.

Additional visualizations are displayed in Fig. 14. As shown, TinyGroundingGPT reveals more distinct location attributions, indicating that it effectively learned multi-scale fine-grained knowledge and achieved high-level alignments among object texts, coordinates, and images. This provides insights for further explaining MLLMs, particularly in grounding tasks. We also provide a demo for utilizing TinyGroundingGPT in Fig. 15.

Model	Quality	Richness	Accuracy
GPT-4V	4.24	4.10	4.88
GroundingGPT-7B	3.68	3.20	3.38
Qwen2-VL-2B	3.90	3.64	4.18
TinyGroundingGPT-3B	4.04	3.90	3.66

Table 13: The assessment for image annotation by GPT-4V includes "Quality" for overall quality, "Richness" for the diversity of object descriptions, and "Accuracy" for precision. Scores are based on the average ratings (1-5) from 50 samples.

Evaluate the image description based on the following criteria:		
Quality (1-5):	**Richness (1-5):**	**Accuracy (1-5):**
1 - The description is incoherent, lacks flow, and does not effectively convey the contents of the image.	1 - The description only mentions a few basic objects or elements in the image, without any contextual details or relationships.	1 - The description contains multiple significant inaccuracies or errors in identifying objects, elements, or their characteristics.
2 - The description has some coherence but is still disjointed, with limited flow and incomplete coverage of the image.	2 - The description includes some additional details about the objects or elements but lacks depth in terms of their relationships or broader context.	2 - The description has some inaccuracies or errors in identifying objects, elements, or their characteristics.
3 - The description is generally coherent, with reasonable flow, and covers most of the key elements in the image.	3 - The description provides a reasonable level of detail about the objects and elements, as well as some of their relationships or broader context.	3 - The description is generally accurate in identifying the objects, elements, and their characteristics, with only minor inaccuracies.
4 - The description is coherent, with good flow, and comprehensively covers the important aspects of the image.	4 - The description is rich in detail, covering a diverse range of objects, elements, their relationships, and the broader context of the scene.	4 - The description is highly accurate in identifying the objects, elements, and their characteristics, with minimal to no inaccuracies.
5 - The description is highly coherent, with excellent flow, and articulately captures the essence of the image in a compelling manner.	5 - The description is exceptionally rich, providing abundant details about the diverse array of objects, elements, their intricate relationships, and the comprehensive context of the scene.	5 - The description is completely accurate in identifying all the objects, elements, and their characteristics, with no discernible errors or hallucinations.

Table 12: The prompt for GPT-4V to assess descriptions from the perspectives of Quality, Richness, and Accuracy.

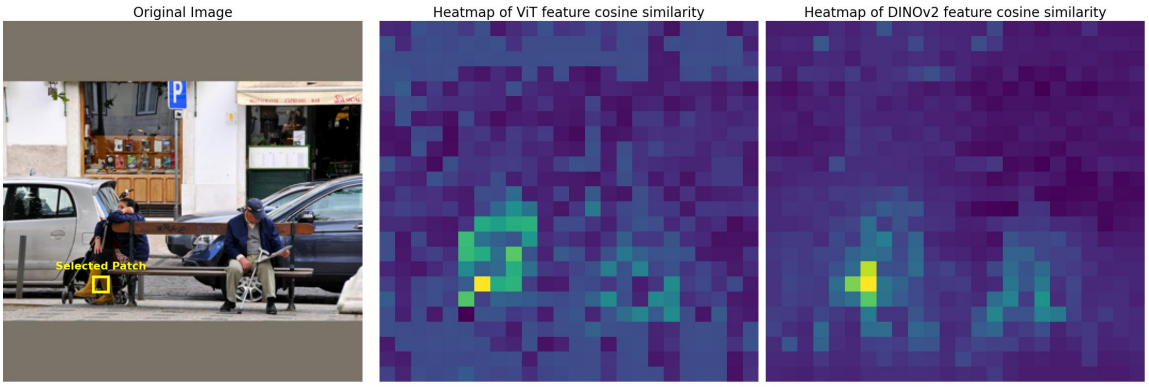


Figure 6: The visualizations of image feature cosine similarity for the selected patch (highlighted with yellow box). Lighter colors indicate higher feature similarity.

Prompt Message for Object Recognition

You are an expert in image recognition, adept at identifying all objects within an image. Please provide a comprehensive list of the objects present, using concise and precise descriptions. Ensure that each entry accurately reflects what is visible, and only identify objects you are certain about. Format your output as follows: [object1, object2, ...].

Prompt Message for Relation Extraction

You are an image relation extraction assistant. Given a list of instances: {Object Recognition}, please provide all relevant potential triples based on the image. Each instance must be selected from the list. The format for the triples is (instance1, relationship, instance2). Output the results as a list, such as [(instance1, relationship, instance2), ...]. Only the list should be provided; do not include any additional text.

Figure 7: The prompt message for object recognition and relation extraction.

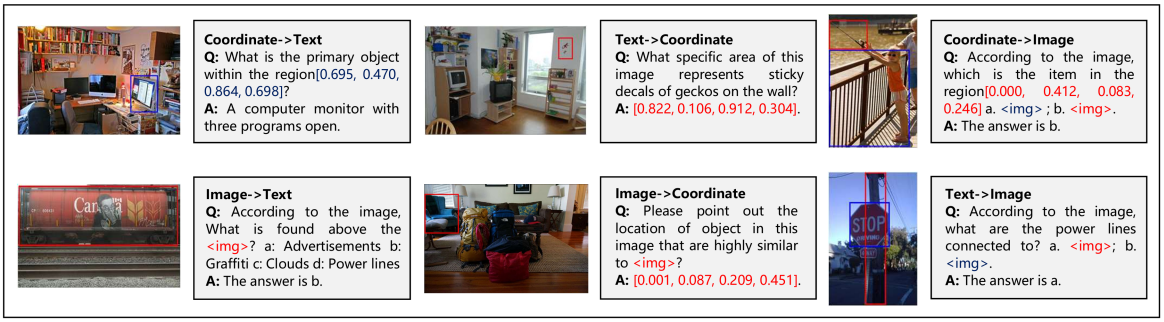


Figure 8: Our generated various kinds of data used in Stage2 for achieving high-level alignments among texts, coordinates, and images, where the < img > denotes the corresponding augmented object image.

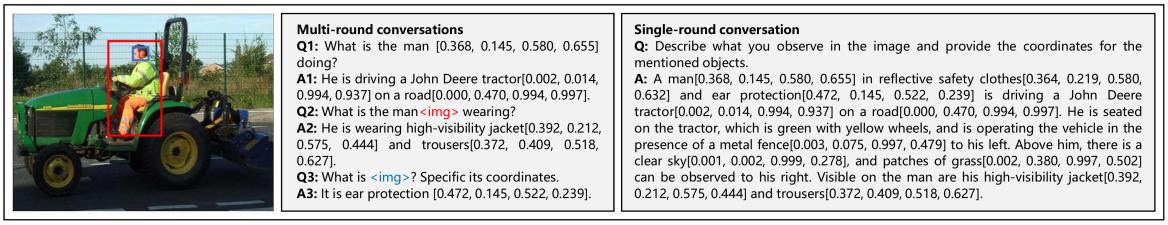


Figure 9: Our generated various kinds of data used in Stage3 for achieving global object alignment, where the < img > denotes the corresponding augmented object image.

Prompt Message

1. You are a QA pair generator assistant. I will provide you with an image: {image}, a list of objects (including the objects present in the image): {Object Recognition}, and a list of triples describing the relationships between the objects : {Relation Extractions}. You may include additional objects and relationships that you are confident about. Please generate 5 optimal QA pairs that accurately reflect the image. Each question must include one object in the list highlighted with "<>", and the options should represent the relationships. Output the QA pairs and corresponding answers in the format of a JSON list. Follow the output format as a JSON list, such as [QA1, QA2, ...]. Do not include any additional text.
2. You are a QA pair generator assistant. I will provide you with an image: {image}, a list of objects (including the objects present in the image): {Object Recognition}, and a list of triples describing the relationships between the objects : {Relation Extractions}. You may include additional objects and relationships that you are confident about. Please generate 5 optimal QA pairs that accurately reflect the image. For each question, the options must consist of only two objects in the list, highlighted with "<>". Output the QA pairs and corresponding answers in the format of a JSON list. Follow the output format as a JSON list, such as [QA1, QA2, ...]. Do not include any additional text.

User Inputs

Object Recognition: A list of objects in the image.
Relation Extractions: A triple list of relationship among objects.
Object Grounding: A dictionary for bounding box coordinates of each subject.

Assistant

Example1
Q1: Can you identity <blue hat> is worn by whom? A. the left woman B. the right man
A1: The answer is B. the right man.
Example2
Q2: What is the right man wearing blue jacket reading? A. <newspaper> B. <cane>
A2: The answer is A. newspaper.

Postprocess

1. Generate text and coordinate alignment QA pairs: (a) Given the object coordinates, formulate questions such as, "Describe the object located within the region [x1, y1, x2, y2]." Provide the corresponding object description as the answer. (b) Given the object description, create questions like, "In this image, where is <description> located?" Provide the corresponding object coordinates as the answer.
2. Generate image and coordinate alignment QA pairs by utilizing object coordinates, such as: "Q: What is the location of <object>? A: <coordinates>" and "Q: Which is the item in region<coordinates>? a. <object1>; b. <object2>. A: The answer is a." Extract the corresponding object images based on the object coordinates and replace <object> with the augmented object images.
3. Generate image and text alignment QA pairs: (a) Identify all objects highlighted with "<>" in the questions or options. (b) Use the dictionary that maps each object to its bounding box coordinates in the image to extract the corresponding object images. (c) Replace the highlighted object in the text with the augmented object image.
4. Ensure that the object images in each QA pair share a small Intersection over Union.

Figure 10: The prompt message and user's input example used for generating our Fine-grained Grounding Dataset in Stage2.

Prompt Message

You are a QA pair generator assistant. I will provide you with an image: {image}, a list of potential object (including objects that may be present in the image): {Object Recognition} , a dictionary of bounding box coordinates (in the format [x1, y1, x2, y2], with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y) for objects: {Object Grounding}, and several sentences that describe the image: {image Annotation}

Based on these materials, the task is to create at least 5 question-answer pairs related to the image based on the information by follow requirements blow:

(1) The object that has bounding box coordinates in the QA pairs should be highlighted with "<>" .

(2) You need include bounding box coordinates after the object that you are certain about in the description in the format "<object> [x1, y1, x2, y2]" instead of directly say the object name or describing the regions in text, ensuring they are as precise as possible.

(3) Avoid introducing objects that do not exist in the original descriptions and refrain from including excessive subjective perceptions to prevent creating illusions.

(4) The boxes provided in different sentences may have some coordinates that are the same or very close, which could be because different expressions refer to the same object. You should analyze and avoid describing a single object within a bounding box as multiple distinct entities

(5) The output should be a list of QA pairs, each of which includes question and answer, and should be limited to 50 words. Keep it as concise as possible.

User Inputs

Origin Annotation (provided in the open source or generated in advance): The description on the image or several sentences describing the sentence.

Object Recognition: A list of objects in the image.

Object Grounding: A dictionary for bounding box coordinates of each subject.

Assistant

{ 'question' : 'What clothes is <the girl>[0.019, 0.302, 0.285, 0.678] wearing?' ,
'answer' : '<The girl> is wearing <a pink dress>[0.019, 0.380, 0.227, 0.596].' }

{ 'question' : 'Why is this <girl> climbing the stairs?' ,
'answer' :' She is going to enter her <playhouse>[0.069, 0.014, 0.997, 0.774].' }

Postprocess

1. Generate more QA pairs about alignment by utilizing object coordinates, such as: "Q: What is <a little girl> ? Specific the location? A: It' s a little girl[0.019, 0.302, 0.285, 0.678] " .
2. Identity the object highlighted with "<>" and coordinates [x1, y1, x2, y2] in the QA pairs, and extract the corresponding object images according to bounding box coordinates.
3. Replace some <object> or [x1, y1, x2, y2] with the augmented object image.

Figure 11: The prompt message and user's input example used for generating our Multi-round Grounding Conversation Data in Stage3.

Prompt Message

You are an image description annotation assistant. I will provide you with an image: {image}, a list of potential instances (including objects that may be present in the image): {Object Recognition} , a dictionary of bounding box coordinates (in the format [x1, y1, x2, y2], with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y) for objects: {Object Grounding}, a relationship triple list (in the format (instance1, relation, instance2)) among objects: {Relation Extractions}, and original image description materials: {Image Annotation}.

Based on these materials, please provide the an accurate and detailed image description by following requirements blow:

(1) You need include bounding box coordinates after the object that you are certain about in the description in the format "object [x1, y1, x2, y2]" instead of directly say the object name or describing the regions in text, ensuring they are as precise as possible.
(2) Avoid introducing objects that do not exist in the original descriptions and refrain from including excessive subjective perceptions to prevent creating illusions.
(3) Be aware that provided instances and their corresponding coordinates may be the same or very close, which could indicate that various expressions refer to the same object. Analyze these cases and avoid describing a single object within a bounding box as multiple distinct entities.
(4) Your output should consist solely of the improved image description, ensuring coherence, accurate, concise, and detailed. The image annotation should be limited to 150 words.

User Inputs

Image Annotation (provided in the open source or generated in advance): The description on the image or several sentences describing the sentence.
Object Recognition: A list of objects in the image.
Relation Extractions: A triple list of relationship among objects.
Object Grounding: A dictionary for bounding box coordinates of each subject.

Assistant

A child[0.019, 0.302, 0.285, 0.678] in a pink dress[0.019, 0.38, 0.227, 0.596] is climbing up a set of stairs[0.0, 0.636, 0.304, 0.958] in an entry way[0.099, 0.052, 0.288, 0.614]. The child is going into her playhouse[0.069, 0.014, 0.997, 0.774], which appears to be a wooden playhouse[0.069, 0.014, 0.997, 0.774]. The playhouse has wooden walls and a decorative sculpture attached next to the entry. There are some flowers around the playhouse, adding a touch of natural beauty.

Figure 12: The prompt message and user's input example used for generating our Grounding Description Data in Stage3.

User: Please provide a detailed description.

Assistant1: our method
Assistant2: general method

Assistant1: In the image, a man[0.412, 0.0, 0.656, 0.882] is seen ironing clothes[0.432, 0.366, 0.628, 0.816] on the back of a yellow taxi[0.432, 0.168, 0.998, 1.0]. He is wearing a bright yellow shirt[0.432, 0.003, 0.656, 1.0] and seems to be focused on his task amidst the hustle and bustle of the city. The taxi, a common sight in urban environments, is captured in motion, driving down a busy street lined with buildings[0.0, 0.0, 0.998, 0.424]. The man's position on the back of the taxi adds an element of surprise and humor to the scene, as it is not a typical location for ironing clothes. The image encapsulates a moment of everyday life in the city, where unexpected activities can take place.

Assistant2: In the image, a man is standing on the back of a yellow taxi cab, engrossed in his work. He's dressed casually in a yellow shirt and blue shorts, and he's holding a blue towel with both hands. The taxi cab, marked with a "NYC Taxi" sign, is in motion on a city street, adding a dynamic element to the scene. In the background, other taxis can be seen, suggesting a busy urban environment. The perspective of the image is from the side of the taxi cab, giving a clear view of the man and his surroundings. The image captures a moment of everyday life in the city, with the man's work as a taxi driver at the center of it.

Figure 13: A comparison of generated image descriptions between TinyGroundingGPT trained with our method and without it.

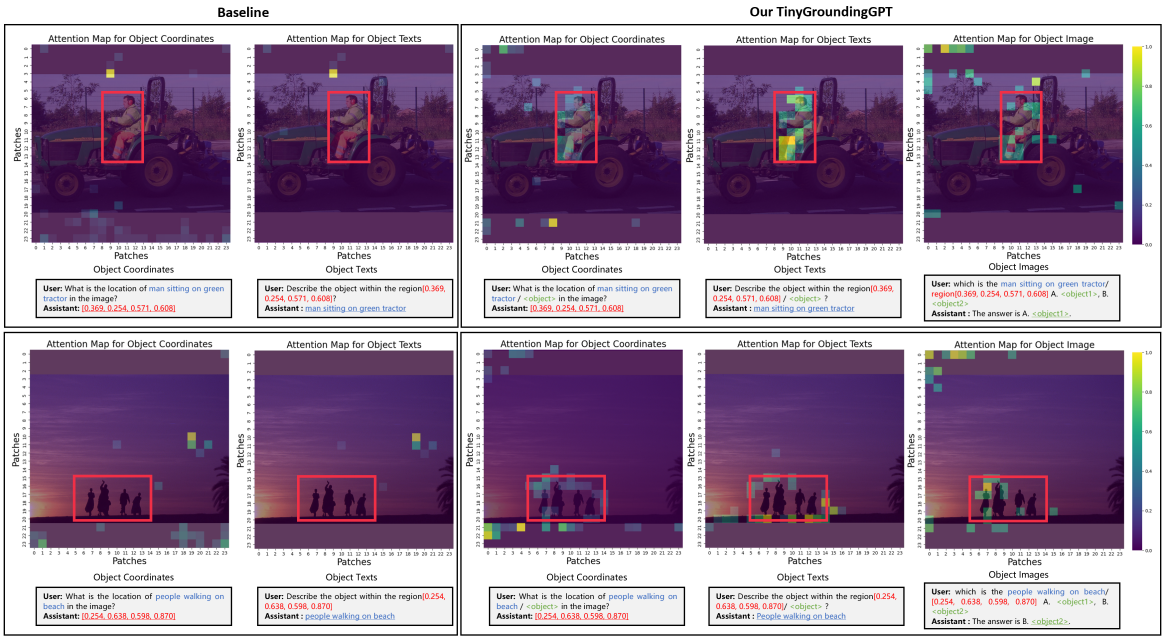


Figure 14: The visualization of the attention map for image patches with different object representation outputs (texts, coordinates, and images, which are underlined), where the red bounding box denotes the target region.

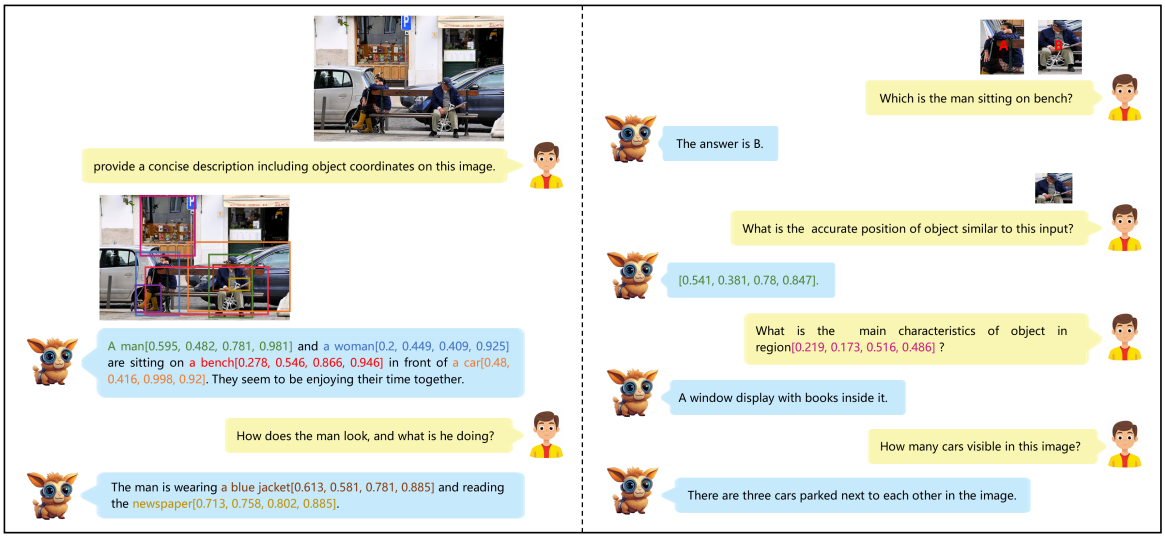


Figure 15: A demo for the use of our TinyGroundingGPT.