Migician: Revealing the Magic of Free-Form Multi-Image Grounding in Multimodal Large Language Models

You Li[1]*, Heyu Huang[2]*, Chi Chen[3]†, Kaiyu Huang[1]†, Chao Huang[1], Zonghao Guo[3],
Zhiyuan Liu[3], Jinan Xu[1], Yuhua Li[2], Ruixuan Li[2], Maosong Sun[3]

[1] State Key Laboratory of Advanced Rail Autonomous Operation,
Beijing Jiaotong University, Beijing, China
[2] Huazhong University of Science and Technology, Wuhan, China
[3] Tsinghua University, Beijing, China

Abstract

The recent advancement of Multimodal Large Language Models (MLLMs) has significantly improved their fine-grained perception of single images and general comprehension across multiple images. However, existing MLLMs still face challenges in achieving precise grounding in complex multi-image scenarios. To address this, we first explore a Chain-of-Thought (CoT) framework that integrates single-image grounding with multi-image comprehension. While partially effective, it remains unstable and struggles to capture abstract visual information due to its non-end-to-end nature. Therefore, we introduce **Migician**, the first multi-image grounding model capable of performing free-form and accurate grounding across multiple images. To support this, we present the MGrounding-630k dataset, which comprises data for several multi-image grounding tasks derived from existing datasets, along with newly generated free-form grounding instruction-following data. Furthermore, we propose MIG-Bench, a comprehensive benchmark specifically designed for evaluating multi-image grounding capabilities. Experimental results demonstrate that our model achieves significantly superior multi-image grounding capabilities, outperforming the best existing MLLMs by 24.94% and even surpassing much larger 70B models. Our code, model, dataset, and benchmark are fully open-sourced at *https://migician-vg.github.io/*.

1 Introduction

Multimodal Large Language Models (MLLMs) have exhibited significant advancements recently, demonstrating exceptional cross-modal understanding capabilities and achieving outstanding performance in various vision-language tasks (Ye et al., 2023; Hu et al., 2024; Elliott and Kádár, 2017; Ive
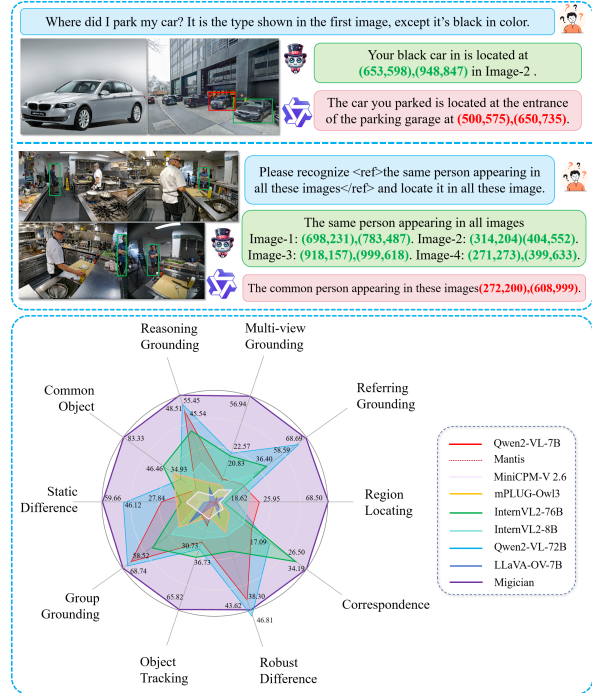

Figure 1: **Top**: Examples of free-form multi-image grounding. The task is to identify and localize relevant visual regions across multiple images based on a free-form query. **Bottom**: Our proposed model, Migician, significantly outperforms other MLLMs on various multi-image grounding tasks.

et al., 2019; Lu et al., 2021; Amini et al., 2019; Krishna et al., 2017). As these models continue to evolve, their capabilities have expanded beyond image-level understanding to include fine-grained visual grounding (Wang et al., 2023; Chen et al., 2023b; You et al., 2023). This enables MLLMs to process region-specific inputs and outputs, unlocking a broader spectrum of real-world multimodal application scenarios (Peng et al., 2023).

Despite the promising visual grounding capabilities demonstrated by existing MLLMs, these abilities are largely confined to single-image scenarios (Kazemzadeh et al., 2014; You et al., 2023). The potential of MLLMs in free-form **multi-image**

---

*Equal contribution.
†Corresponding authors: Kaiyu Huang (kyhuang@bjtu.edu.cn) and Chi Chen (chenchithu@gmail.com).

grounding (MIG) remains underexplored. Free-form MIG challenges the model to perform grounding across multiple images effectively, where the input queries and image contexts can be organized in arbitrary forms, enabling flexible and dynamic interactions. For instance, as shown in Figure 1, the model must understand the white car in the query image and relate it to the textual prompt "black in color" to identify the corresponding target in the target image. This capability unlocks a wide range of applications, such as fine-grained environmental perception in autonomous driving (Wang et al., 2024c), anomaly detection in surveillance systems (Black et al., 2002), and target localization for embodied robotics (Grauman et al., 2022). To address the free-form MIG, the model needs to possess the capability for visual grounding while achieving cross-image understanding.

As a result, a question naturally arises: *Can we integrate the single-image grounding and multi-image understanding capabilities of existing MLLMs to tackle the MIG task?* In this work, we propose a Chain-of-Thought (CoT) framework that first leverages multi-image understanding to generate a textual referring query, and then utilizes it for localization through single-image grounding. This approach is proven highly effective for MIG tasks, particularly in simple scenarios where textual descriptions are sufficiently distinctive, demonstrating the potential of MLLMs in handling such tasks.

However, the proposed CoT framework struggles with describing abstract visual semantics in multi-image scenarios, and the two-step process results in a doubling of the inference time. To address this, we further propose Migician, a competitive MLLM capable of free-form and accurate grounding across multiple images, serving as an end-to-end solution for MIG. To progressively establish flexible grounding capabilities, we employ a two-stage training procedure based on our proposed large-scale MIG dataset (MGrounding-630k). First, the grounding ability of Migician is enhanced through a combination of data from MIG tasks and general tasks. Then, Migician is further refined using high-quality free-form MIG instruction data. In addition, to evaluate the challenges of the free-form MIG scenario, we construct a comprehensive multi-image grounding benchmark, MIG-bench, comprising a total of 10 different tasks, 5.9k diverse images and more than 4.2k test instances. We observe a significant gap between the performance of existing mainstream MLLMs and human

performance on the MIG-bench. In contrast, Migician can effectively alleviate this gap and improve the performance of free-form MIG.

To sum up, our contributions can be concluded as follows:

- We explore the task of multi-image grounding for MLLMs and reveal the potential and challenges of current MLLMs by through a proposed CoT framework.

- We introduce Migician, the first MLLM capable of effectively performing free-form MIG. We also present MGrounding-630k, the first large-scale MIG instruction tuning dataset for training this model.

- We introduce MIG-Bench, a comprehensive benchmark for evaluating multi-image grounding capabilities. Experimental results demonstrate that Migician significantly outperforms the current best methods.

2 Related Work

Multimodal Large Language Models Recent developments in multimodal large language models (MLLMs) have shifted from single image-text understanding towards more versatile capabilities (Cai et al., 2024; Yao et al., 2024; Wang et al., 2024b; Li et al., 2024a). Among these efforts, some focus on enabling models to achieve fine-grained visual grounding, either through simple instruction tuning (Chen et al., 2023b; Peng et al., 2023) or by integrating additional auxiliary visual components (You et al., 2023; Zhang et al., 2023; Chen et al., 2023a). However, these models primarily focus on visual grounding within a single image. Some other studies explore multi-image understanding tasks, such as multi-image comparison, reasoning, and temporal comprehension (Jiang et al., 2024; Li et al., 2024c; Ye et al., 2024; Li et al., 2024a; Cai et al., 2024; Yao et al., 2024). Nevertheless, fine-grained visual grounding at the multi-image level remains an underexplored area. To the best of our knowledge, our proposed Migician is the first MLLM designed to address the challenge of multi-image grounding.

MLLM Benchmarks Most existing benchmarks for evaluating MLLMs focus on single-image tasks (Fu et al., 2023; Li et al., 2024b). A few

Figure 2: An illustration of the multi-image grounding tasks included in MIG-Bench. These tasks are divided into two categories: spontaneous grounding and referential grounding, depending on the whether there are explicit referential requirements.

recent benchmarks have started assessing the performance of MLLMs on multi-image understanding (Jiang et al., 2024; Meng et al., 2024; Fu et al., 2025; Wang et al., 2024a; Liu et al., 2024a), but they primarily emphasize image-level comprehension. The most relevant benchmark to our work is MC-Bench (Xu et al., 2024), a contemporaneous study. MC-Bench evaluates the multi-context grounding capabilities of MLLMs by asking them to accurately locate the corresponding object based on a text prompt in the correct image from a given pair. However, it exhibits limitations in the fixed number of input images and the restricted forms of queries. In contrast, the proposed MIG-Bench in this work offers more flexible task formats, focusing on evaluating models' capabilities in free-form multi-image understanding.

3 Task Definition

The task of free-form multi-image grounding is to identify and localize relevant visual regions across a set of images based on a free-form query. Unlike traditional grounding tasks with fixed input formats, the query in free-form multi-image grounding can be an **arbitrary combination of text and images**, making it highly flexible and versatile. Formally, let the query Q consist of a natural language description, reference images $\{R_1, R_2, \ldots, R_k\}$ or a hybrid combination of both (e.g., "[a white car image] find a car like this image except it is black"). Given a set of target images $\{I_1, I_2, \ldots, I_n\}$, the task is to identify a set of visual regions $\{G_1, G_2, \ldots, G_m\}$ where G_i is a region within an image I_j that satisfies the semantic and contextual constraints defined by Q.
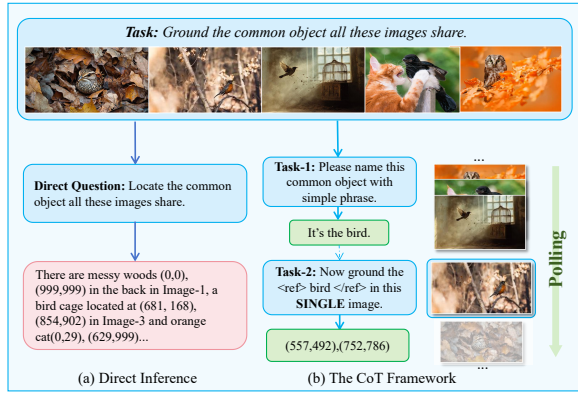
Figure 3: Illustration of the CoT framework. Different from (a) direct inference, (b) the CoT method decomposes the task into two subtasks, solving each task deploying the model's existing capabilities. Polling refers to queryi the MLLM with one image at a time.

As shown in Figure 2, based on whether the task involves explicit reference requirements, multi-image grounding tasks can be further categorized into two types: *Spontaneous Grounding* and *Referential Grounding*. Spontaneous Grounding refers to recognizing and grounding the target object in corresponding images without explicitly pointing it out. Unlike the conventional Reference Expression Comprehension task (Kazemzadeh et al., 2014) that explicitly refer to the target object, Spontaneous Grounding typically utilizes the relationships between multiple images as contextual cues to autonomously identify and localize the objects to be grounded (e.g., finding and locating differences between images). Referential Grounding, on the other hand, requires an explicit reference to the target object. As mentioned earlier, such references can take the form of arbitrary combinations of images and textual descriptions.

4 Methods

In this section, we delve into methods for enabling free-form multi-image grounding capabilities in MLLMs. We begin by exploring a Chain-of-Thought (CoT) framework to elicit the capabilities within existing MLLMs to tackle this task. Then we develop an end-to-end MIG model, Migician, through instruction tuning to overcome the limitations of the CoT framework and achieve enhanced MIG performance.

4.1 A Chain-of-Thought Framework

As illustrated in Figure 3(a), directly prompting existing MLLMs to perform MIG tasks often leads to

significant performance degradation. To better explore the potential of existing models, we design a CoT framework which decouples the MIG task into two stages as shown in Figure 3(b). The model first understands the input images and question, generating a textual referring expression that describes the target object. Next, the model locates the objects in corresponding images using the referring expression from the previous step.

This framework leads to a notable performance improvement on MIG tasks. However, the CoT framework has inherent limitations, such as error propagation due to its multi-step process, which also reduces reasoning efficiency (Yao et al., 2022). Additionally, many scenarios require grounding through abstract visual semantics across multiple images, which cannot be effectively captured through textual expressions. More failure types are detailed in Appendix C. This highlights the need for an end-to-end model capable of directly performing the MIG task.

4.2 Data Construction

The CoT framework has demonstrated that an MLLM with both multi-image understanding and single-image grounding capabilities inherently holds strong potential for free-form MIG. In the following section, we employ instruction tuning to explicitly bridge these capabilities in existing MLLMs to achieve MIG. For this purpose, we first construct an instruction tuning dataset for MIG, named **MGrounding-630k**, with its statistics presented in Figure 4. This dataset is primarily constructed through the following two ways.

Transforming Existing Data. By analyzing the tasks and annotation types of existing datasets, we identify multiple multi-image grounding (MIG) tasks whose data could be derived through transformation of the existing. Specifically, we collect and organize data from existing sources, combining or automatically synthesizing single-image annotations to create datasets for 6 types of MIG tasks. Each task contains over 70k examples, resulting in a total of 530k training samples. The details of these task data can be found in Appendix D.1.

Synthesizing Free-form MIG Data. The data obtained through the aforementioned methods still do not fully meet the requirements for free-form MIG. To acquire MIG data with richer and more diverse formats, which would enhancing the model's
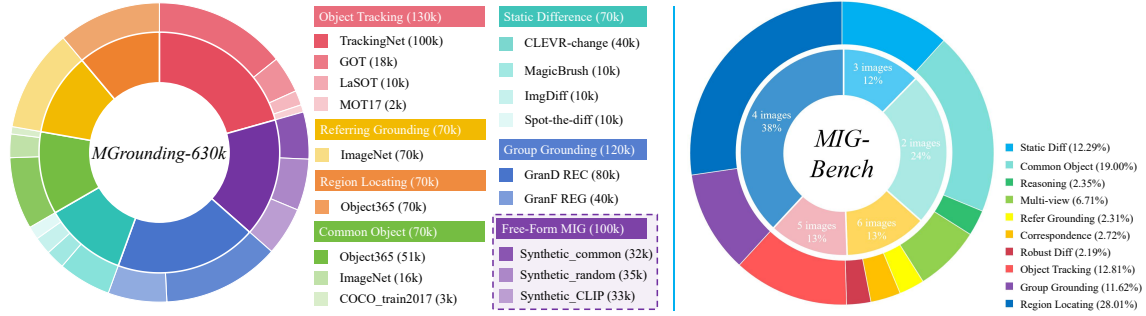
Figure 4: Statistics of the MGrounding-630k dataset and MIG-Bench.

instruction-following and flexible grounding capabilities, we design a MIG data synthesis pipeline. This pipeline uses the Object365 (Shao et al., 2019) images with object annotations, select multiple images as a group, and generate high-quality instructions for multi-image grounding. Specifically, we first employ Qwen2-VL-72B (Wang et al., 2024b) to generate captions of each individual image and then perform error filtering and refinement on the annotated bounding boxes. Next, we prompt Qwen2.5-72B (Yang et al., 2024) to automatically generate high-quality, free-form MIG question-answering pairs by integrating information from multiple images. To optimize the selection of appropriate image groups, we adopt different image grouping methods, including random selection, selection of images with common objects, and grouping images based on CLIP similarity to select semantically similar images for each. Using these methods, we generate a total of 100k Free-Form MIG data. Detailed information can be found in Appendix D.2.

4.3 Instruction Tuning for MIG

Using the constructed dataset, we perform instruction tuning based on Qwen2-VL-7B (Wang et al., 2024b) to develop **Migician**, enabling it to achieve end-to-end free-form MIG capabilities.

Two-Stage Training. To effectively equip the model with free-form MIG capabilities, we propose a two-stage training approach. In the first stage, the model learns to perform multi-image grounding by training on the six representative MIG tasks of MGrounding-630k, acquiring the ability to simultaneously comprehend multiple images and execute visual grounding. In the second stage, the model is further fine-tuned on free-form MIG instruction data in MGrounding-630k, enabling it to adapt to more flexible and diverse instruction types and

transfer the MIG skills learned in the first stage to a broader range of scenarios. To prevent the model from forgetting its existing capabilities during training, we also incorporate single-image understanding, multi-image understanding, and single-image grounding data into each training stage. For more details please refer to the Appendix E.

Model Merging. After the second stage of fine-tuning, we observe a trade-off between model performance and flexibility: while the model adapts to the free-form MIG instructions, there is a performance drop in common multi-image grounding tasks. To better balance these two aspects, we adopt the model merging technique (Ilharco et al., 2022), averaging the model weights obtained from both training stages as the final weights. We find this approach mitigates the performance loss in common MIG tasks while preserving the ability to follow free-form MIG instructions effectively.

5 MIG-Bench

We introduce MIG-Bench, a manually curated benchmark designed to evaluate the MIG ability of current MLLMs. It comprises 5.9k images, and 4.3k testing instances, covering 10 distinct tasks shown in Figure 2 with details in Appendix A. The distribution of these tasks is illustrated in Figure 4.

MIG-Bench is manually constructed from multiple data sources. Initially, we select annotated data examples from existing datasets and adapt them for the MIG task. We collect challenging examples from Objects365 (Shao et al., 2019) for Common Object Grounding and Region Locating. We use examples that exhibits significant movement from GOT-10k (Huang et al., 2019) for Object Tracking. For Multi-View Grounding, we utilize the rich annotations from Ego4D (Grauman et al., 2022). The Static Difference task is sourced from MagicBrush (Zhang et al., 2024). We combine multi-

Models	Spontaneous Grounding			Referential Grounding							AVE
	Difference		Similarity	Visual Reference				Textual	Visual+Textual		
	Static	Robust	Common	OT	MV	Region	Refer	GG	Reason	Co-Re	
Human Performance											
Human	99.50*	97.87	98.00*	100.00	96.88	100.00*	98.99	91.06*	92.08	97.44	97.18
70B-Scale MLLMs											
LLaVA-OV-72B	13.26	5.34	26.84	12.91	7.64	2.14	17.83	21.60	11.88	8.55	13.65
InternVL2-76B	15.91	10.64	36.40	30.73	20.83	5.74	46.46	41.28	32.67	26.50	26.72
Qwen2-VL-72B	46.12	46.81	64.46	26.73	22.57	18.62	33.33	62.53	50.50	17.09	38.88
7B-Scale MLLMs											
Shikra	0.00	2.13	0.74	0.00	0.00	0.83	1.01	0.60	0.00	0.85	0.62
GroundingGPT	0.38	0.00	19.85	0.18	0.35	1.75	2.02	19.04	5.94	0.00	4.95
Ferret	8.90	7.45	14.09	2.36	0.69	1.41	2.02	33.27	12.87	1.71	8.48
Mantis	1.52	0.00	3.31	12.18	2.08	1.00	1.01	10.02	0.00	0.85	3.20
LLaVA-OV-7B	6.06	3.19	3.43	0.18	1.04	1.08	9.09	15.43	6.93	0.85	4.73
MiniCPM-V 2.6	14.58	2.13	14.34	9.82	6.25	1.75	11.11	10.02	2.97	2.56	7.55
mPLUG-Owl3	18.56	6.38	34.93	8.55	7.64	2.41	7.07	22.85	9.09	5.98	12.35
InternVL2-8B	6.92	7.45	25.49	20.73	9.72	3.49	28.28	30.26	17.82	9.40	15.96
Qwen2-VL-7B	27.84	38.30	19.36	20.73	11.81	25.95	23.23	58.52	48.51	11.97	28.62
mPLUG-Owl3$_{+CoT}$	16.29	8.51	55.39	44.36	25.35	19.04	36.36	30.86	18.81	10.26	26.52
InternVL2-8B$_{+CoT}$	14.58	7.45	72.54	40.91	27.78	28.60	67.68	44.49	41.58	11.97	35.76
Qwen2-VL-7B$_{+CoT}$	23.48	40.43	63.85	62.73	42.71	24.85	54.55	43.29	51.49	30.77	43.82
Migician	**65.15**	**46.81**	**84.19**	60.07	70.73	66.53	74.31	76.77	59.41	**34.19**	**63.82**

Table 1: Performance comparison of different models on MIG-Bench. OT, MV, GG and Co-Re respectively means object tracking, multi-view grounding, group grounding and correspondence. For values marked with *, we randomly sample 20% testing examples for human evaluation on the corresponding task.

ple examples from GranD (Rasheed et al., 2024) to form a group for Group Grounding. Additionally, for tasks such as Reasoning Grounding, Correspondence, Referring Grounding, and Robust Diff Grounding, which lack suitable existing datasets, we collect both web images and manually captured photos, and annotate them with well-educated annotators who are thoroughly trained and fully understand the requirements of these tasks.

All instances are reviewed by two different human annotators to guarantee the quality of MIG-Bench. This includes removing instances with incorrect annotations, ensuring that the questions are answerable, filtering out overly simplistic questions, and refining ambiguous queries. We also invite five volunteers to answer the questions to evaluate human performance on this benchmark (detailed in Appendix F). As shown in Table 1, the average accuracy of human responses is 97.18%, indicating that the task is easy for humans and further demonstrating the high quality of MIG-Bench.

Unlike existing benchmarks, MIG-Bench introduces the grounding task in a multi-image setting, thereby addressing the gap in previous benchmarks that are unable to measure the free-form MIG ca-

pabilities of MLLMs. A detailed comparison with other benchmarks is provided in Appendix B.

6 Experiments

6.1 Implementation Details

Migician undergoes development based on the Qwen2-VL-7B (Wang et al., 2024b) foundation model with a global batch size of 48, a total of 25,000 steps for the two-stage training procedure, and a learning rate of 5e-6, using 8×A100-80G GPUs. For the evaluation in our proposed MIG-Bench, we use the conventional metric Acc$_{0.5}$ in referring expression comprehension (Kazemzadeh et al., 2014). This metric measures the accuracy of object localization, defining a prediction as correct if the Intersection over Union (IoU) with the ground truth bounding box is greater than 0.5.

6.2 Results on MIG-Bench

As shown in Table 1, Migican achieves the state-of-the-art performance across all tasks on MIG-bench, with an average improvement of 24.94% compared to the second-best model, Qwen2-VL-72B (38.88%), despite having significantly fewer parameters. Note that there is a substantial gap be-

Model	MuirBench	BLINK val	MIBench	Mantis_eval	MMIU	AVE
Closed-Source Model						
GPT-4o	62.31	60.04	71.88	62.67	55.7	62.52
Gemini-Pro	49.35	45.16	—	—	53.4	49.30
Open-Source Model						
LLaVA-1.5	23.46	37.13	26.83	31.34	19.20	27.59
CogVLM	20.85	41.54	—	45.16	23.57	32.78
Idefics2-8B	26.08	—	46.39	48.85	27.80	37.28
mPLUG-Owl3	39.67	50.30	56.66	63.10	21.72	46.29
InternVL2-8B	48.70	50.57	52.91	60.37	42.00	50.05
Mantis	_44.50_	49.05	45.09	57.14	45.60	48.28
LLaVA-OV-7B	41.80	48.20	_71.29_	64.20	44.46	53.99
MiniCPM-V 2.6	42.65	51.45	71.09	_69.12_	50.19	56.90
Qwen2-VL-7B	42.04	**52.35**	68.06	**70.97**	_54.36_	_57.56_
Migician	**53.69**	_51.53_	**71.42**	_69.12_	**60.32**	**61.51**

Table 2: Performance comparison on various multi-image understanding benchmarks. The highest score is highlighted in bold and the second highest score is *underlined* for all open-source models.

Model	MME	MMBench	V* Bench
Closed-Source Model			
GPT-4V	1926.6	_81.0_	54.97
Gemini-Pro	2148.9	73.6	48.16
Claude-3.5	1920.0	79.7	—
Open-Source Model			
LLaVA-1.5	1510.7	64.3	48.68
InternVL2-8B	_2210.3_	**81.7**	43.07
MiniCPM-V 2.6	2024.6	77.2	52.67
SEAL	1128.9	33.1	**75.39**
LLaVA-OV-7B	1998.0	80.9	—
Mantis	1806.4	75.7	—
Migician	**2244.7**	80.0	_72.30_

Table 3: The performance of models on various single-image benchmarks, where Migician consistently exhibits strong capabilities.

Models	Spontaneous	Referential	AVE
mPLUG-Owl3	19.96	9.08	13.04
mPLUG-Owl3$_{+mCoT}$	23.78	14.10	17.62
mPLUG-Owl3$_{+CoT}$	26.73	26.43	26.54
InternVL2-8B	13.29	17.10	15.71
InternVL2-8B$_{+mCoT}$	23.78	21.99	22.64
InternVL2-8B$_{+CoT}$	31.52	37.57	35.37
Qwen2-VL-7B	19.96	28.67	28.61
Qwen2-VL-7B$_{+mCoT}$	41.83	26.23	31.90
Qwen2-VL-7B$_{+CoT}$	42.59	44.34	43.70

Table 4: The comparison among different CoT variants. We compare three representative MLLMs among direct reference, single-image CoT (*+CoT*), multi-image CoT (*+mCoT*) as described in Section 7.1.

tween human performance and that of all MLLMs across all tasks, indicating that MLLMs have significant potential for improvement in free-form MIG. In particular, for 7B-scale models, even advanced multi-image models like InternVL2-8B and Qwen2-VL-7B struggle to perform, particularly in tasks such as multi-view grounding, region locating, and correspondence.

For models equipped with grounding capabilities, such as mPLUG-Owl3, InternVL2 series, and Qwen2-VL series, they demonstrate an advantage over other baselines. Furthermore, the proposed single-image CoT method (*+CoT*) effectively integrates the grounding and multi-image understanding capabilities of the MLLMs where different abilities assist each other in different reasoning steps, achieving comprehensive improvements on multi-image grounding tasks. Moreover, this approach proves effective for all the aforementioned models.

Traditional single-image visual grounding models like Shikra (Chen et al., 2023b), Grounding-

GPT (Li et al., 2024d) and Ferret (You et al., 2023) do not inherently support multi-image input, generally demonstrating poor performances in MIG-Bench. Their subpar performance can be attributed to two key factors: (1) These grounding MLLMs are typically designed for single-image grounding and are generally incapable of effectively processing direct multi-image input or concatenated images. (2) These models lack essential capabilities for following instructions of free-form multi-image grounding. In comparison, Migician achieves significantly better performance on MIG tasks, effectively extending the capability range of grounding MLLMs.

6.3 Results on Multi-Image Understanding Benchmarks

As shown in Table 2, Migician not only establishes its multi-image grounding ability, but also remarkably stimulates its general multi-image understanding ability. In particular, Migician achieves the best average results on the multi-image understanding benchmarks. It surpasses the second-best model (Mantis) on MuirBench by 9.19%, achieving SOTA performance on both MMIU and MIBench. We attribute this to the training on a mixture of multi-image understanding and grounding data, which indicates that our proposed MGrounding-630k dataset can effectively enhance general multi-image comprehension.

6.4 Results on Single-Image Benchmarks

Table 3 lists the empirical results on typical single-image understanding benchmarks including MME (Fu et al., 2024) and MMBench (Liu et al.,

Setting	Multi-image General Benchmarks					MIG
	MuirBench	BLINK	MIBench	Mantis	MMIU	
Base	42.04	52.35	68.06	70.97	54.36	28.62
Full data	53.77	51.27	71.76	66.36	53.31	62.79
-w/o grounding	$44.54_{(-9.23)}$	$51.32_{(+0.42)}$	$71.68_{(-0.08)}$	$67.74_{(+1.38)}$	$52.12_{(-1.19)}$	$22.43_{(-40.36)}$
-w/o general	$53.62_{(-0.15)}$	$49.25_{(-2.02)}$	$65.22_{(-6.54)}$	$64.52_{(-1.84)}$	$48.61_{(-4.70)}$	$62.21_{(-0.58)}$

Table 5: The ablation study on the contribution of different data subsets.

Model	Easy	Medium	Hard
# Instances	2471	1430	395
InternVL2-8B	44.69	13.92	1.77
Qwen2-VL-7B	30.31	22.24	0.00
InternVL2-8B$_{+CoT}$	67.10	7.06	0.25
Qwen2-VL-7B$_{+CoT}$	71.02	10.70	0.76
Migician	**76.00**	**52.10**	**29.37**

Table 6: Performance comparison across varying difficulty levels on MIG-Bench.

2024b), suggesting that Migician retains strong single-image understanding capacities. Specifically, on the MME benchmark, Migician surpasses notable MLLMs like InternVL2-8B and MiniCPM, while exhibiting equally strong performance with close-source models on MMBench. Notably, contrasting with specialized multi-image models such as LLaVA-OV and Mantis, whose single image ability has largely degenerated, Migician poses significant advantage over them, achieving comprehensive capability maintenance. Our model also maintains strong performance on single-image REC tasks, as shown in Figure 8.

Furthermore, we find that the MIG ability of Migician can be leveraged to address the task of finding visual details in high-resolution images, such as V*Bench (Wu and Xie, 2024). Specifically, we split a single high-resolution image in V*Bench into multiple sub-images and transform the problem into a MIG task (detailed in Appendix F). Results show that Migician can generalize well to this out-of-distribution setting, performing on par with the specialized visual searching system SEAL (Wu and Xie, 2024).

7 Analysis

7.1 Effects of Different CoT Strategies

The CoT framework in Section 4.1, after obtaining a referring expression, has the MLLM perform grounding in each image in a polling manner (denoted as single-image CoT), which incurs significant inference overhead. Here, we explore multi-image CoT, where the MLLM directly performs grounding across all images based on the obtained referring expression. As shown in Table 4, multi-image CoT achieves some effectiveness but it still falls significantly behind single-image CoT. In contrast, our proposed Migician is able to perform end-to-end reasoning, offering significant advantages in both efficiency and effectiveness.

7.2 Effects of Different Data on Multi-Image Understanding

As observed in Table 2, Migician shows an improvement in multi-image understanding. We further conduct an ablation study to analyze the effects of different data subsets. Specifically, we train two models with either MGrounding-630k or multi-image understanding data removed from the training set. The results in Table 5 reveal that grounding data generally aids multi-image understanding. In 4 out of 5 benchmarks, the full dataset achieves the highest performance compared to models trained with any subset of data removed. In contrast, directly fine-tuning with only general data does not consistently lead to a performance boost. However, when combined with fine-grained grounding data, the model experiences a notable improvement.

7.3 Performance across Difficulty Levels

To comprehensively assess Migician across varying scenario complexities in MIG-Bench, we have established a three-tier difficulty classification (Easy/Medium/Hard) through joint consideration of three key factors: (1) the number of input images, (2) the accuracy rates of four representative baselines (Qwen2-VL, InternVL2, and their CoT variants), and (3) the average IoU improvement when applying CoT. Specifically, an instance is classified as **Easy** if either (a) more than two models achieve

correct answers with fewer than four input images, or (b) the CoT-enhanced models demonstrate an IoU improvement exceeding 0.15. Conversely, instances are deemed **Hard** when no more than one model succeeds despite processing over four input images. All remaining cases that fall between these thresholds are categorized as **Medium** difficulty.

Through this way, we can evaluate model capabilities across diverse scenario complexities. As shown in Table 6, Migician demonstrates remarkable performance advantages across all difficulty levels while exhibiting particularly strong capabilities in challenging scenarios. Notably, the performance gap widens significantly in medium and hard difficulty settings as Migician achieves approximately 30% greater accuracy compared to baseline models. This pronounced advantage in complex scenarios highlights Migician's superior capacity for free-form MIG.

8 Conclusion

In this work, we explore the task of multi-image grounding and propose Migician, the first MLLM to overcome the barriers between fine-grained visual grounding and multi-image inputs. With our proposed large-scale MGrounding-630k dataset, Migician seamlessly integrates grounding across multiple images, enabling free-form multi-image grounding. To further advance research in this area, we introduce MIG-Bench, a comprehensive benchmark for evaluating the multi-image grounding capabilities of MLLMs. Experimental results demonstrate that our model significantly outperforms existing methods. We hope this work will inspire further developments in multi-image grounding and contribute to the creation of more versatile multimodal models in the future.

Limitation

Despite our comprehensive discussion of the MIG challenge, several limitations remain. First, due to computational constraints, we have not verified the effectiveness of our training methods on larger 70B-scale models. Second, the current model can still produce hallucinated outputs sometimes as conventional MLLMs. Lastly, our training methods and benchmark construction mainly focus on the REC task. Although Migician possesses decent REG capacity, this topic is still insufficiently discussed.

References

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*.

James Black, Tim Ellis, and Paul Rosin. 2002. Multi view image surveillance and tracking. In *Workshop on Motion and Video Computing, 2002. Proceedings.*, pages 169–174. IEEE.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. 2023a. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Desmond Elliott and Akos Kádár. 2017. Imagination improves multimodal translation. *arXiv preprint arXiv:1705.04350*.

Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. 2019. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

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*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2025. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*.

Lianghua Huang, Xin Zhao, and Kaiqi Huang. 2019. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. *arXiv preprint arXiv:1906.07701*.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Qirui Jiao, Daoyuan Chen, Yilun Huang, Yaliang Li, and Ying Shen. 2024. Img-diff: Contrastive data synthesis for multimodal large language models. *arXiv preprint arXiv:2408.04594*.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024b. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024c. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

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

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

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, et al. 2024a. Mibench: Evaluating multimodal large language models over multiple images. *arXiv preprint arXiv:2407.15272*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi

Wang, Conghui He, Ziwei Liu, et al. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. 2024. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*.

Anton Milan. 2016. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*.

Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4624–4633.

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*.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

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

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

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

Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. 2023. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. 2024c. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759.

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*.

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

Yunqiu Xu, Linchao Zhu, and Yi Yang. 2024. Mcbench: A benchmark for multi-context visual grounding in the era of mllms. *arXiv preprint arXiv:2410.12332*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

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*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. 2023. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

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*.

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

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2024. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36.

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

A Benchmark Tasks Definition

A.1 Spontaneous Grounding

Our benchmark evaluates spontaneous grounding through three distinct tasks below, which aim at assessing model's ability to autonomously discover insidious connections across various images and then accurately locate the target.

Spot the Difference Similar to the spot-the-difference puzzle, given two similar images with a single difference, the model is instructed to recognize and ground this difference in the second image, requiring simultaneous and keen perception of both images.

Common Object Grounding It refers to automatically recognizing and grounding the common object appearing in all images within an image group, which shares one definite common object in our benchmark.

Robust Image Difference Grounding Models must focus on the primary difference between two images captured from slightly different perspectives, ignoring other minor variations caused by shifts in the viewpoint. The incorporation of view changes presents a greater challenge for the model and better reflects real-world scenarios, where variations in perspective are inevitable.

A.2 Reference Grounding

Textual Reference Query This challenge, which mainly includes **Group Grounding**, tests a model's ability to link a textual reference to the target object within its corresponding image among an image group. Given a set of images and one textual query, the model must firstly identify the correct image then accurately ground the target object within it, additionally incorporating image-level locating compared with conventional grounding task.

Visual Reference Query These tasks focuses on effectively utilizing visual reference information and incorporate it into the locating process.
(1) **Visual Referring Grounding.** In this task, a pair of images is provided—a source image with a clear object and a target image containing multiple elements. The model must perceive the referenced object of the source image and then locate it in the target image.
(2) **Region Locating.** Models are tasked with identifying multiple region images within a source image, which often requires perceptive and discerning

observation as the model may encounter person recognition, similar object distinguishing, tiny item searching and etc.
(3) **Object Tracking.** This task involves tracking a target object across a sequence of video frames. The object is highlighted with a red bounding box in the first image, and the model must keep track of it throughout the entire sequence.
(4) **Multi-view Grounding.** Here, the model must locate the same target across multiple images taken from distinctive viewpoints, for instance egocentric view and exocentric view.

Visual+Textual Reference Query These tasks combine information from both modalities to assess cross-modal reasoning&grounding abilities.
(1) **Correspondence.** The model must ground semantically or functionally similar regions within the target image. This finer-grained task focuses on object regions rather than whole objects, demanding an in-depth understanding of visual semantics.
(2) **Reasoning.** This task requires the model to perform reasoning-based grounding by integrating cross-modality information. Several examples are shown in Figure 2.
Our MIG-Bench offers a rich, multi-faceted evaluation across various real-world scenarios and domains, extending beyond simple image pairs to include longer and more complex image contexts. By ensuring that each task is well-defined and unambiguous, we facilitate objective and definitive assessments.

B Comparison of MIG-Bench with Other Benchmarks

A detailed comparison with other benchmarks are provided in Table 7. Current evaluations for MLLMs primarily focus on single-image perception, understanding, reasoning, or grounding (e.g., MME, MMBench, Refcoco), leaving the multi-image scenario largely unexplored. While benchmarks such as Mantis-eval, BLINK, and MIRB are representative of multi-image evaluation, they concentrate on image-level and general understanding of multiple images, failing to comprehensively assess a model's fine-grained grounding skills in the multi-image scenario.

Unlike traditional grounding benchmarks or multi-image benchmarks, MIG-Bench introduces the grounding task into a multi-image scenario, covering a series of 10 distinct tasks. The most relevant benchmark to our work is MC-Bench, a

Dataset	Images	Ave-I	Max-I	Multi-Image	Multi-Task	Instance-Labeled	Instances	Reference
Q-Bench	3489	2.0	2.0	✘	✔	✘	✘	✘
Mantis-Eval	542	2.5	5.0	✔	✘	✘	✘	✘
BLINK	3612	1.9	4.0	✔	✔	✘	✘	✘
MIRB	3497	3.8	42.0	✔	✔	✘	✘	✘
Refcoco/g/+	3900	1.0	1.0	✘	✘	✔	7596	T
HC-Refcoco/+/g	1521	1.0	1.0	✘	✘	✔	3754	T
GigaGround	3775	1.0	1.0	✘	✘	✔	61353	T
MC-Bench	3345	2.0	2.0	✘	✔	✔	3202	T
MIG-Bench	5887	3.8	6.0	✔	✔	✔	4295	T/I/T+I

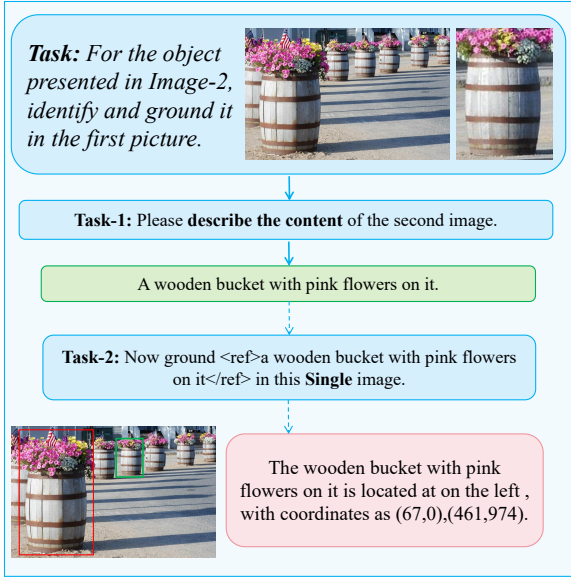Table 7: Comparison of MIG-Bench with other benchmarks.

Figure 5: A CoT framework failure case about abstract and intricate visual semantics.

contemporaneous study. MC-Bench evaluates the multi-context grounding capabilities of MLLMs by asking them to accurately locate the corresponding object based on a text prompt in the correct image from a given image pair. However, it has limitations in terms of the fixed number of input images and the restricted forms of queries. In contrast, the proposed MIG-Bench offers more flexible task formats, focusing on evaluating models' abilities in free-form multi-image understanding.

C Single-Image CoT Failure Patterns

Despite the significant effectiveness of our proposed CoT framework, its limitation remains, particularly when dealing with visually intricate cases, as shown in Figure 5. For example, a detailed textual reference such as "A wooden bucket with pink flowers on it" is insufficient for adequately representing or identifying the target object.

Furthermore, we have analyzed more failure patterns of the CoT framework in Figure 6, categorized into perceptual and reasoning flaws.

For the former, the framework falls short when multiple images are organized in a manner where only integrating all their visual information could tackle MIG (i.e. finding the location of missing people in the second image), or when the textual content could not sufficiently represent the visual information.

Regarding reasoning errors, inaccuracies can arise at various stages of the reasoning process, undermining the framework's overall accuracy and effectiveness.

These failure patterns highlight the significant limitations of simply integrating the different capabilities of current models through a simple CoT framework, underscoring the need for an end-to-end model capable of directly performing the MIG task.

D MGrounding-630k Data Curation Details

D.1 Transforming Existing Data

Static Diff Describing the differences among two nearly identical pictures is a well discussed topic, yet previous attempts capture the differences through textual phrases, failing to precisely recognize their locations. After a comprehensive survey on this area, we have collected high-quality and fully labeled image difference data from various existing datasets: Spot-the-diff (Jhamtani and Berg-Kirkpatrick, 2018), Img-diff (Jiao et al., 2024), MagicBrush (Zhang et al., 2024) and CLEVR-change (Park et al., 2019).

By these collected datasets inherently contain much noise in them, for instance, inaccurate difference caption, incorrect bounding box coordinates
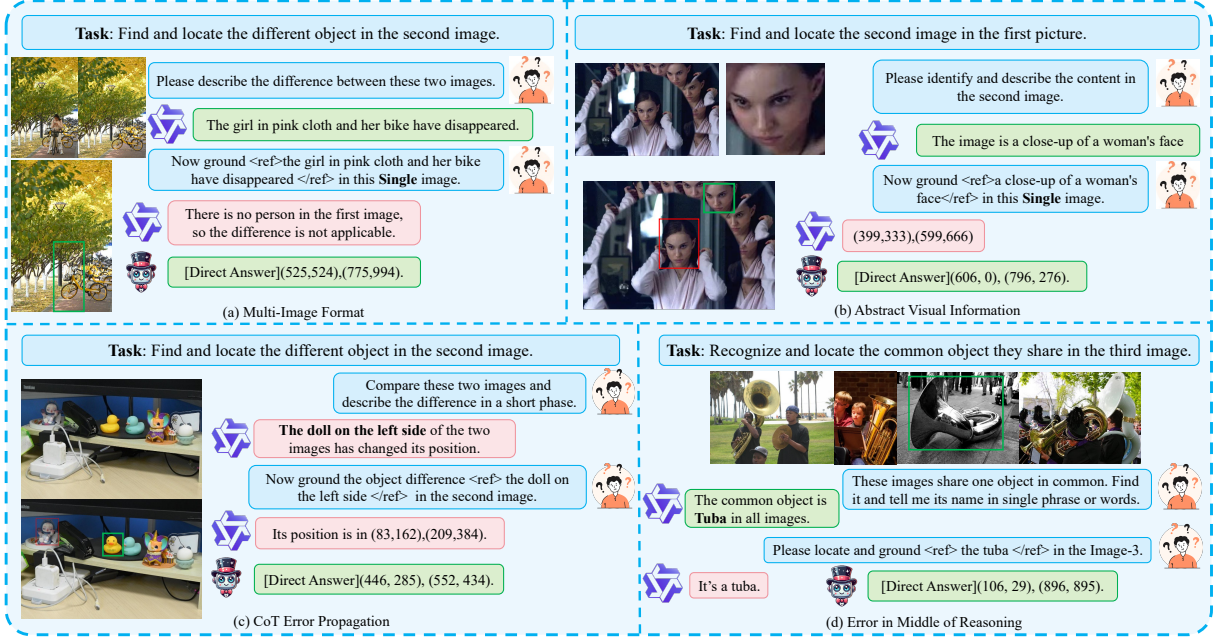
Figure 6: Above are the four representative failure patterns of the single-image CoT. From left to right, top to bottom, they are (a) special multi-image format, (b) abstract visual information, (c) CoT error propagation, (d) step-2 inference error.

Model	RefCOCO			RefCOCO+			RefCOCOg		AVE
	val	testA	testB	val	testA	testB	val	test	
VisionLLM v2 (Wu et al., 2024)	79.20	82.30	77.00	68.90	75.80	61.80	73.30	74.80	74.14
Shikra (Chen et al., 2023b)	87.00	90.60	80.20	81.60	87.40	72.10	82.30	82.20	82.97
InternVL2-8B (Cai et al., 2024)	87.10	91.10	80.70	79.80	87.90	71.40	82.70	82.70	82.94
GroundingGPT (Li et al., 2024d)	88.02	91.55	82.47	81.61	87.18	73.18	81.67	81.99	83.57
Griffon v2 (Zhan et al., 2024)	89.60	91.80	86.50	81.90	85.50	76.20	85.00	86.00	85.30
InternVL2-8B (Cai et al., 2024)	87.10	91.10	80.70	79.80	87.90	71.40	82.70	82.70	82.94
Qwen2-VL-7B (Wang et al., 2024b)	**91.70**	**93.60**	**87.30**	85.80	90.50	79.50	87.30	**87.80**	87.96
Migician	91.62	93.49	87.22	**86.13**	**91.06**	79.93	**88.06**	87.80	**88.16**

Table 8: Performance of Models on single image grounding benchmark.

and etc. We filter the inaccurate bbox labels from Spot-the-Diff and only preserve the correct ones and their difference captions. Additionally, we conduct down-sampling on Img-Diff due to its diffusion generation (Rombach et al., 2022) based nature and consequent inaccuracies.

During the construction process, we ensure the diversity of the content by (1) incorporating numerous prompt formats generated by GPT-4 and improving the instruction-following ability of the model, (2) constructing CoT process to assist the model gradually and progressively reaching the final answer by fully utilizing the annotation available in the dataset.

Common Object Grounding Grounding the primary common object across multiple images is a challenging task for models. It requires them to simultaneously perceive multiple images, isolate the common object, and then accurately ground the target within each image.

In our approach, we leverage diverse data sources, including ImageNet (Deng et al., 2009), COCO (Lin et al., 2014), and Object365, which are rich in extensive annotations. To organize the images, we group those containing the same object based on object labels, applying a threshold to filter out objects that occupy too small a proportion of the image. This threshold, determined empirically, effectively mitigates ambiguity when multiple candidate objects could be considered the common

Training Methods	Referring	Object Tracking	Group Grounding	Region	Static Diff	Common Object
Base	23.23	20.73	58.52	25.95	27.84	19.36
Multi-Task Learning	60.00	61.65	62.28	57.95	55.68	81.37
Separate Learning	69.70	74.55	63.13	65.42	68.94	79.53
Model Merging	60.61	50.00	64.53	18.95	29.92	65.44

Table 9: Comparison between different training methods. We compare the learning efficiency between multi-task learning, separate learning and merging all these task-specialized modes. We mainly focus on the in-domain tasks that M-Grounding dataset covers.

Models	Settings	Common Object	Multi-view Grounding	Object Tracking	Region Locating
—	Random Guess	26.47	1.04	2.13	0.00
Qwen2-VL-7B	Polling	19.96	11.83	20.73	25.95
Qwen2-VL-7B	All	19.36	6.60	13.09	11.80
Qwen2-VL-7B	Polling+mCoT	53.80	14.24	21.09	20.20
Qwen2-VL-7B	All+mCoT	45.71	9.38	17.55	15.54
Migician	Polling	81.99	44.44	61.09	59.65
Migician	All	72.43	43.06	58.55	34.91

Table 10: Comparison of different answering forms. For random guess, we set the default answer as *(0,0),(999,999)*.

object, thereby producing clear and definitive training examples. We further reduce ambiguity by removing classes that often co-occur with other object, for instance keyboard, knife, couch, dinning table and etc.

Object Tracking The multi-image setting is well-suited for the object tracking task, which involves both temporal and spatial grounding. To build our dataset, we have carefully selected large-scale, well-annotated datasets including TrackingNet (Muller et al., 2018), LaSOT (Fan et al., 2019), GOT-10K (Huang et al., 2019), and MOT-2017 (Milan, 2016). During dataset construction, we extract 4-6 images from each original sequence at specific intervals, preserving key features while maintaining efficiency. Additionally, we introduce a small proportion of temporal order judgment data from continuous video frames to strengthen the model's temporal reasoning capabilities.

Referring Grounding This training data simulates the process of locating an object from a source image in a target image. We primarily use the ImageNet-2012 dataset to create image pairs, where the source object is fully visible in the first image, and the target object is hidden in the second. Empirically, this design generates challenging training examples that require the model to first recognize the source object and then search for it in the target image.

Group Grounding Conventional visual grounding is typically limited to single-image contexts. However, in real-world scenarios, it is often necessary to recognize a target object within a cluttered collection of images. Group Grounding addresses this limitation by enabling the identification of the target among a group of images, thereby enhancing the versatility of traditional grounding methods.

For the construction of Group Grounding training data, we leverage the large-scale GranD rec and reg conversation dataset (Rasheed et al., 2024), which contains 3 million examples. After filtering out noisy data and grouping 3-5 images per set, we curate a high-quality collection of 120k instances for stage-1 training. This dataset effectively enhances the model's ability to perform image co-reference (Jiang et al., 2024), image-level grounding, and instance-level localization.

Region Locating Region locating involves slicing an image into several semantically rich regions and identifying the precise locations of these regions within the source image. To extract meaningful regions, we use the Objects365 dataset, selecting labeled bounding box areas as the regions of interest. To further enhance the quality of the regions, we apply a series of filtering criteria: (1) Content Richness: We select images with more than 10 bounding box annotations to avoid overly simple or plain cases. (2) Aspect Ratio: We retain regions with an aspect ratio between 0.5 and 2 to exclude excessively narrow bounding boxes that may

Algorithm 1 CLIP Adaptive Similarity Selection

Require: Images $\mathbf{I}$, adaptive selection range k, $thres \in (0,1)$
Ensure: Final Image Set $\mathbf{F}$
1: Initialize $\mathbf{F} \leftarrow \emptyset$
2: Extract $\mathbf{F}_I \leftarrow$ Features of $\mathbf{I}$
3: **while** $\mathbf{F}_I$ is not empty **do**
4:     Randomly select $thres \sim \text{Uniform}(0.1, 1)$
5:     **for** each $f_i \in \mathbf{F}_I$ **do**
6:         $s_{ij} = \text{similarity}(f_i, f_j), \forall f_j \in \mathbf{F}_I, j \neq i$
7:     **end for**
8:     $\mathbf{Sort\_S}_i = \text{Sort}(s_{ij})[1:]$
9:     $k \leftarrow \lfloor thres \times (\text{len}(\mathbf{Sort\_S}_i)) \rfloor$
10:     $\mathbf{Candidates} \leftarrow \mathbf{Sort\_S}_i[:k]$
11:     Randomly select $r \sim \text{Uniform}(3, 5)$
12:     $\mathbf{Selected} \leftarrow \text{Sample}(\mathbf{Candidates}, r)$
13:     Append f_i and $\mathbf{Selected}$ to $\mathbf{F}$
14:     Remove f_i and $\mathbf{Selected}$ from $\mathbf{F}_I$
15: **end while**
16: **return** $\mathbf{F}$

---

be difficult for the model to handle. (3) Size: We ensure that the region-to-image ratio lies between 0.2 and 0.49, with an absolute pixel count above 2,000, to exclude tiny and obscure regions that may lack sufficient detail. Notably, due to our carefully designed filtering mechanism and the inherent characteristics of the task, the resulting training data predominantly includes cases of person recognition, distinguishing between similar objects (i.e. chairs, bowls, cars and etc), and recognizing tiny details—tasks that are non-trivial even for humans.

D.2 Synthesizing Free-form MIG Data

The algorithm for CLIP adaptive similarity image input is shown in Algorithm 1. We further display our prompt template for image caption generation, bounding box label refinement and instruction tuning data generation in the following pages and several stage-2 data examples in Figure 9.

Specifically, we deploy Qwen2-VL-7B for detailed image caption generation and Qwen2-VL-72B for bbox label refinement. The inference process is accelerated through vLLM framework (Kwon et al., 2023).

E Details of Two-Stage Training

This section outlines the data proportions and their respective sources for the two training stages, as summarized in Table 12.

In stage 1, we leverage both single-image and multi-image datasets encompassing general understanding and grounding tasks to comprehensively enhance the model's capabilities. At this stage, the stage-1 subset from MGrounding-630k constitutes the largest portion of the training data, with a total of 530k examples. The total training examples for stage-1 is 1 million.

In stage 2, the focus shifts to stimulating the model's free-form MIG abilities by integrating all free-form grounding data from MGrounding-630k. A significant proportion of stage-1 data is also reused to maintain the previously learned abilities. The total number of training examples in this stage is 200k.

F Evaluation Implementation

Polling-based Evaluation When directly requiring the model to generate bounding box coordinates for each image, due to their limited multi-image grounding ability and insufficient instruction following ability, the answer obtained in this way is largely unfaithful and mostly unsatisfactory in instruction following, failing to objectively reflecting the real grounding ability of the model. . Empirically, instructing the models to directly generate all bounding box coordinates results in very serious instruction following issue. The models struggle with plausible outputs, with their answers mostly containing only one bounding box or pure text analysis.

Considering current model's feeble performance, we transform from directly generating all answers to polling every single image, which facilitates definite and objective evaluation. Empirically, directly generating all the bounding box coordinates for all images results in lower performance. Yet as illustrated in Table 10, Migician still demonstrates great robustness to the variation of evaluation format.

V*Bench Evaluation Implementation We transform the high-resolution single image grounding task into a MIG challenge. Specifically, we slice a single high-resolution image into multiple sub-images and directly convert the problem into the group grounding task, which first requires the model to perform image-level locating and then ground the target in that specific image. By utilizing the MIG ability of Migician, we can locate the regions relevant to the input question. Afterward, the model combines the identified region with the original image to generate the answer for the input

Models	Spontaneous Grounding			Referential Grounding							AVE
	Difference		Similarity	Visual Reference				Textual	Visual+Textual		
	Static	Robust	Common	OT	MV	Region	Refer	GG	Reason	Co-Re	
70B Scale Models											
LLaVA-OV-72B	13.26	5.34	26.84	12.91	7.64	2.14	17.83	21.60	11.88	8.55	13.65
InternVL2-76B	15.91	10.64	36.40	30.73	20.83	5.74	46.46	41.28	32.67	26.50	26.72
Qwen2-VL-72B	**46.12**	46.81	64.46	26.73	22.57	18.62	33.33	62.53	50.50	17.09	38.88
LLaVA-OV-72B$_{+CoT}$	20.27	21.28	52.57	44.36	20.83	25.60	37.37	35.07	31.68	28.21	31.72
InternVL2-76B$_{+CoT}$	16.86	6.38	**70.34**	**70.55**	33.33	27.27	68.69	57.31	52.48	23.08	42.63
Qwen2-VL-72B$_{+CoT}$	33.33	**47.87**	69.24	70.18	**60.42**	**51.04**	**78.79**	**70.74**	**70.30**	**35.04**	**58.70**

Table 11: Performance Comparison of 70B scale models equipped with CoT.

Type	Source	Ratio
Stage-1		
S-Understanding	LLaVA-OV-data	17%
S-Grounding	RefCOCO series, Groma-Instruct	13%
M-Understanding	M4-Instruct(Li et al., 2024c)	16%
M-Grounding	MGrounding-630k (Stage-1)	54%
Stage-2		
S-Understanding	LLaVA-OV-data	9%
S-Grounding	RefCOCO series, Groma-Instruct	7%
M-Understanding	M4-Instruct(Li et al., 2024c)	8%
M-Grounding	M-Grounding (Stage-1)	27%
	M-Grounding (Stage-2)	49%

Table 12: Training data proportion for two stages.

question, achieving high-accuracy results.

70B Scale Models The performances of three competitive 70B scale models are illustrated in Table 11 when equipped with single-image CoT. The general effectiveness of CoT framework is tremendous, with the average performance boost at 20 points. Yet even competitive and much larger model like Qwen2-VL-72B (58.70%) still can't surpass our Migician (60.49%) in multi-image grounding, demonstrating great competence.

Single-Image Grounding As presented in Table 4, Migician not only acquires free-form multi-image grounding capabilities but also demonstrates continual and consistent performance improvements on the RefCOCO series single-image grounding benchmark, surpassing specialized grounding models such as Griffon v2 and GroundingGPT by a large margin. Additionally, Migician outperforms Qwen2-VL-7B in terms of average scores.

Human-Level Evaluation We engaged five human volunteers to answer questions from MIG-Bench. For the static image difference, common object grounding, region locating, and group ground-

ing tasks, we randomly selected 20% of the test examples for efficient evaluation.

The volunteers are instructed to answer the MIG questions by directly drawing bounding boxes on the images, with the resulting annotations then compared to the ground truth using automatic IoU calculation. To ensure unbiased results, evaluations from different volunteers were conducted separately, preventing mutual distraction or information leakage.

G Multi-Task Learning

Our whole training process involves the learning process of multiple distinct tasks. How does the actual learning efficiency alter compared with learning these tasks separately, can they contribute to each other or comprise to some extent?

We conduct experiments that only expose the model to omni-task dataset and the results are shown in Table 9. It clearly reveals the conflicts of learning various tasks, with mixes multi-task training consistently surpassing omni-task learning by a huge margin. When we directly merge the checkpoints of all these trained specialized models (Ilharco et al., 2022), the merged model fail at excelling at most tasks, with the average performance falling behind simple multi-task learning.

H Case Study

We provide detailed cases comprehensively reflecting the free-form MIG ability of Migician in Figure 7, 8, as well as our instruction tuning data details examples in Figure 9.

Prompt Template for Caption and Instruction Data Generation

Bbox Refinement Template
Now I'd like you to inspect the original image carefully. Then filter, refine and enhance these annotated objects. Finally, just give me your final modified annotations.

Filtering
Based on you insightful observation of the image, please eliminate the obviously inaccurate (object,bbox) pairs, which in supposed to be small in quantity.

Refine
Refine and enhance the original class/name of each object into a short yet richer caption containing its attributes like color, position, feature(e.g plane <|box_start|>(x1,y1),(x2,y2)<|box_end|> -> dark gray plane flying in the sky <|box_start|>(x1,y1),(x2,y2)<|box_end|>).

Amplify
If any important objects are missing from the annotations, and you believe they are significant and essential, and you are confident of their location, feel free to add them to the final annotations.

Output Format
Modified object caption followed by its bounding box coordinates.

Now the original bounding box annotations I give to you are:

Caption Generation
Describe this image thoroughly in a fluent paragraph. Include all the objects and their attributes(color, shape, size and feature), relative position and relationship.

Multi-image Grounding Instruction Generation

Template 1
Based on the following detailed information of multiple images, please compose meaningful and flexible CROSS-IMAGE grounding questions that link different objects across the images by their attributes similarity/contrast—such as color, position, features, gender, size, shape, etc.—or by other potential logical connection between them.
Specifically:
1.The questions should include CROSS-IMAGE grounding requests that requires the answer to identify and locate various potentially connected object across different images. You can use the connection or similarity between these objects to refer the target item.
2.When referring an object in the question, keep the reference description concise and avoid giving away unnecessary information(like bbox or over-detailed caption) that could lead to answering too easily. You are encouraged to refer the target object to be grounded by the connection of these objects, instead of explicitly point out the object. For instance: "ground the car in image-2 that contrasts most in quality with the shabby vehicle in image-4", rather than "ground the fancy red sports car(explicitly pointing out) in image-2 that contrasts most in quality with the shabby vehicle in image-4", by doing so we can also introduce a bit reasoning process.
3.Include the bounding box coordinates of referred object in the answer as well as the explanation. (Actually you can get a lot of information from the coordinates, which are formatted as (x1,y1),(x2,y2))
4.Strictly format the output as simple Q: A:. In answer, follow the format <ref>object</ref> for objects mentioned. Below are the detailed image captions and the objects in the corresponding images:

Template 2
According to the detailed description of each image, the key objects' captions and their corresponding bboxes below, please provide inferential and free-form question-answer pairs around these different fine-grained information for cross-image grounding/locating, by mining the information and correlations between different objects and different images. You can also get a lot of information from the coordinates, which are formatted as (x1,y1),(x2,y2).
Several Question-Answer Examples to better understand my intention(these examples are not necessarily related with the image information below, they are just examples):
Q: There is a group of people walking around the bus in Image-1. There are also many other people in other pictures. Yet in image-3, I'd like to you locate the person of the same gender with the group of people in image-1. Analyze the problem and locate it precisely.
A: The group of people wearing suits walking together are all men. In Image-3, the person of the same gender is <ref>the man reading a book</ref> at (245,784)(456,924). There is also a woman accompanying him by his side, yet not the same gender of the people in Image-1.

Please read these image information carefully and response in strict plain format Q: A:, follow the format <ref>object</ref>(x1,y1),(x2,y2) for objects and bboxes in answer, and avoid revealing overly detailed explanations (including bbox) in Q to make answering too easy.

Template ...

Prompt Template for Single-Image CoT

Task: Static diff
Step-1: Compare these two images carefully and tell me where does they differ. Please answer briefly in single phrase or words.
Step-2: According to the object difference/change: [RESPONCE], please ground this difference with bounding box coordinates.

Task: Robust diff
Step-1: Compare these two images carefully and describe the prominent different object with really simple words or phrase.
Step-2: Now ground the object difference/change : "[RESPONCE]" with bounding box coordinates.

Task: Referring Grounding
Step-1: Watch carefully and briefly describe the object in the Image-1.
Step-2: Please find and ground the object <|object_ref_start|>[RESPONCE]<|object_ref_end|> with bounding box coordinates.

Task: Common Object
Step-1: These images share one object in common. Recognize it and tell me its name in single phrase or words.
Step-2: Please locate and ground the target object according to the reference: <|object_ref_start|> [RESPONCE] <|object_ref_end|>

Task: Region Locating
Step-1: Describe the content of the XXXth picture with simple phrase or words.
Step-2: Please ground the object <|object_ref_start|>[RESPONCE]<|object_ref_end|> with bounding box coordinates.

Task: Multi-View
Step-1: Describe the object in the first image marked with red bounding box(<|box_start|> (A,B),(C,D) <|box_end|>) with simple phrase or word. You can refer to other images for more precise recognition and description.
Step-2: Locate and ground the object <|object_ref_start|> [RESPONCE] <|object_ref_end|> with bounding box coordinates.

Task: Object Tracking
Step-1: Describe the object in the first image marked with red bounding box with simple phrase.
Step-2: Now ground the target moving object [RESPONCE] with bounding box coordinates.

Task: Group Grounding
Step-1: Just recognize and tell me which image is it in. Answer from: Image1 | Image2 | Image3...
Step-2: [Selected Image] + [Original Question]
Note: For group grounding, the single image at step-2 is selected by matching the answer from step-1. If the framework fails to extract the target image, we send the first image by default.

Task: Reasoning
Step-1: [Original Question] + Name this object in the Image-2 with simple phrase.
Step-2: Please locate and ground the object <|object_ref_start|>[RESPONCE]<|object_ref_end|> with bounding box coordinates.

Task: Correspondence
Step-1: For the first image, describe the semantic/functional feature of the area marked by the red bounding box (<|box_start|>(A,B),(C,D)<|box_end|>).
Step-2: Ground the area that shares the same semantic or functional meaning of: [RESPONCE].

Format Prompt
Format: <|box_start|>(x1,y1),(x2,y2)<|box_end|>. Don't generate additional words.
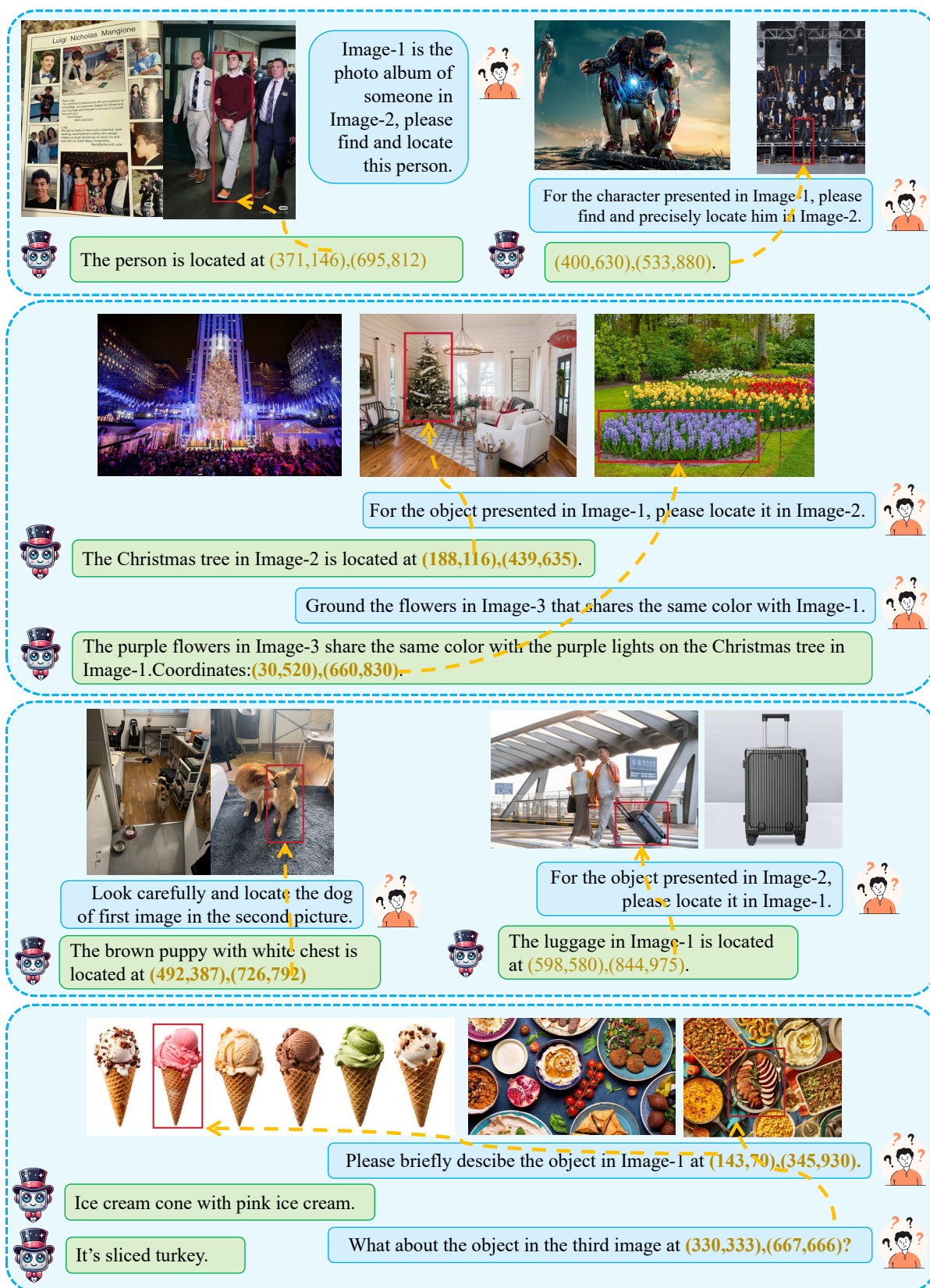**Note: we deploy this prompt for better instruction following.**

Figure 7: Example cases of the free-form multi-image grounding ability of Migician.

For the people in Image-1, at their time, which object on the table of Image-2 may help them travelling around the world?

The globe on the table in Image-2 may help the people in Image-1 traveling around the world. It is located at **(0,208),(208,468)**.

Ground the specific item on Image-1's table that can divide and split the item in Image-2 into pieces.

The black scissors located at **(292,30),(406,806)** could cut the pile of white paper in Image-2.

Focus on three objects: Image-1 (500,500),(700,834), Image-2 (0,0),(579,486), Image-3 (283,500), (612,795). Please briefly describe them.

Image-1: Ferris wheel in the distance.
Image-2: A colorful carousel with people riding on it.
Image-3: A large bed in the center of the room with a white comforter.

Good, now based on all these objects, please compose a detailed and compelling story about our happy trip today.

We had a great time at the amusement park today. We started by enjoying the Ferris wheel, which offered a great view of the park and the surrounding area. Then we moved on to the carousel, where we had a lot of fun. Finally, we go back home to sleep on the cosy bed. Overall, it was a perfect day at the amusement park.

In Image-2, please ground the bottle shares the similar shape with those in image-1, except its color is dark purple.
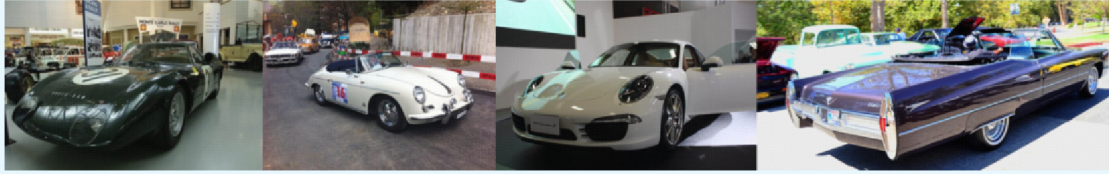
The tall wine glass filled with red wine **(258,100),(470,838)**.

For the object marked with red box in Image-1, please locate in the last image.

The basketabll is located at **(695,58),(743,130)**

Figure 8: Example cases of the free-form multi-image grounding ability of Migician.

Q: In Image-1, the dark-colored classic racing car has a large white circle with the number "1" on the front. Can you find a similar numerical marking on any of the cars in Image-2? If so, where is it located?

A: In Image-2, the vintage white convertible car has a blue and white rally plaque with the number "16" on it. The coordinates for this plaque are (178,302) to (912,776).

Q: Image-3 features a white sports car with a sign on the front bumper reading "911 Carrera S." Is there a similar sign or label on any of the cars in Image-1? If so, what does it say and where is it located?

A: In Image-1, there are no signs or labels on the cars that are explicitly mentioned. However, there are informational signs and banners around the area, including one prominent banner that reads "MONTE CARLO RALLY." The coordinates for the vintage sports car in the background are (0,255) to (466,598), but the specific sign is not detailed in the provided information.

Q: In Image-2, the vintage white convertible car is being driven by a person wearing glasses and a hat. Can you identify a person wearing a hat in Image-4? If so, where is the person located?

A: In Image-4, there is a person wearing a hat located at coordinates (561,213) to (676,330). The hat itself is located at coordinates (565,211) to (607,284).

Q: In Image-1, there is a red bottle on the countertop located at (906,717) to (931,851). Can you find a similar object in Image-3, and if so, provide its bounding box coordinates?

A: In Image-3, there is no red bottle specifically mentioned. However, there is a stack of papers and books on the desk located at (0,753) to (224,932). While it's not a bottle, it is another object that could be considered similar in the context of a workspace.

Q: In Image-4, there is an old, rusted sedan located at (7,285) to (965,700). Can you find a vehicle in Image-2 that contrasts sharply with the state of this car, and provide its bounding box coordinates?

A: In Image-2, there is a large white coach bus with the name "CARADOC" on its side and front destination sign located at (133,186) to (719,821). This bus is clean and well-maintained, contrasting sharply with the neglected and rusted sedan in Image-4.

Q: In Image-2, there is a tall building with a modern architectural style featuring large glass windows and a sleek facade located in the background. Can you find a building in Image-4 that contrasts with this modern style, and provide its bounding box coordinates?

A: In Image-4, there are several houses in the background at (20,208) to (980,330) with various architectural styles, including some with gabled roofs and others with flat roofs. These houses, located in the background, contrast with the modern building in Image-2.

Figure 9: Training Examples of the free-form instruction tuning data.