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

Anonymous ACL submission

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 21.61% and even surpassing much larger 70B models.

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 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

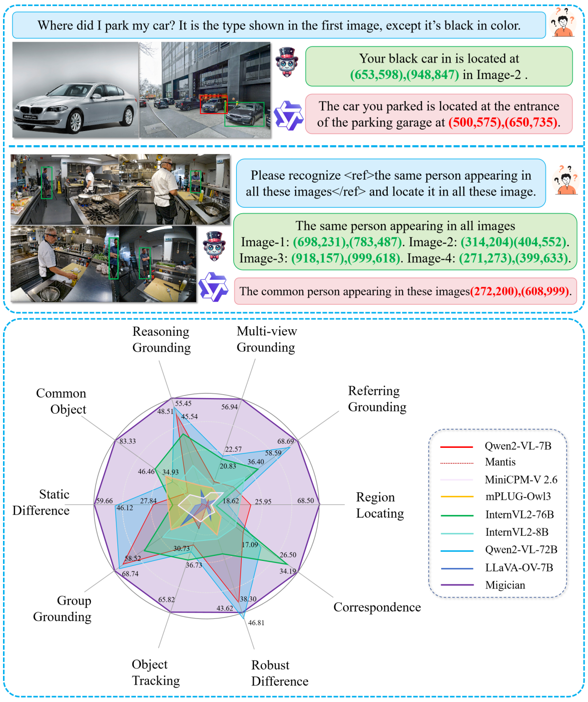


Figure 1: **Top**: Examples of free-form multi-image grounding. The task is to identify and localize relevant visual regions across multiple images based on a free-form query. **Bottom**: Our proposed model, Migician, significantly outperforms other MLLMs on various multi-image grounding tasks.

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 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 integrated 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, which is 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 of 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 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., 2024), but they primarily emphasize image-level comprehension. The most relevant benchmark to our work

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.

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.

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 the methods for enabling free-form multi-image grounding capabilities in MLLMs. Since free-form MIG requires the ability to perform visual grounding while simultaneously understanding multiple images, we begin by investigating a Chain-of-Thought (CoT) framework to combine these two capabilities within existing MLLMs to tackle this task. Furthermore, 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

Although some existing MLLMs such as Qwen2-VL-7B (Wang et al., 2024b) demonstrate strong multi-image understanding and single-image grounding capabilities, we find that directly prompting them to perform MIG tasks often leads to significant performance degradation as illustrated in Figure 3(a). To better explore the potential of existing models for MIG tasks, we design a Chain-of-Thought (CoT) framework that enables the model to effectively leverage and combine its exitsing abilities during the MIG execution.

Specifically, we decompose the MIG task into two subtasks as illustrated in Figure 3(b). The model is first prompted to engage in a "reasoning process" by performing multi-image understanding based on the input images and the given prompt, generating a textual referring expression that describes the target object. Next, the model performs the visual grounding task, using the referring expression from the previous step to locate the objects in corresponding images. This framework leads to a notable performance improvement on MIG tasks, indicating that existing MLLMs possess the underlying capabilities required for such tasks but need an effective method to elicit them.

However, the CoT framework suffers from several inherent limitations. On one hand, the multi-step process introduces error propagation issues (Yao et al., 2022) and impacts reasoning efficiency. On the other hand, many scenarios require grounding through abstract visual semantics across multi-image contexts (as shown in Figure 3(c)), making the use of an intermediate textual referring expression impractical. 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 C.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 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 au-

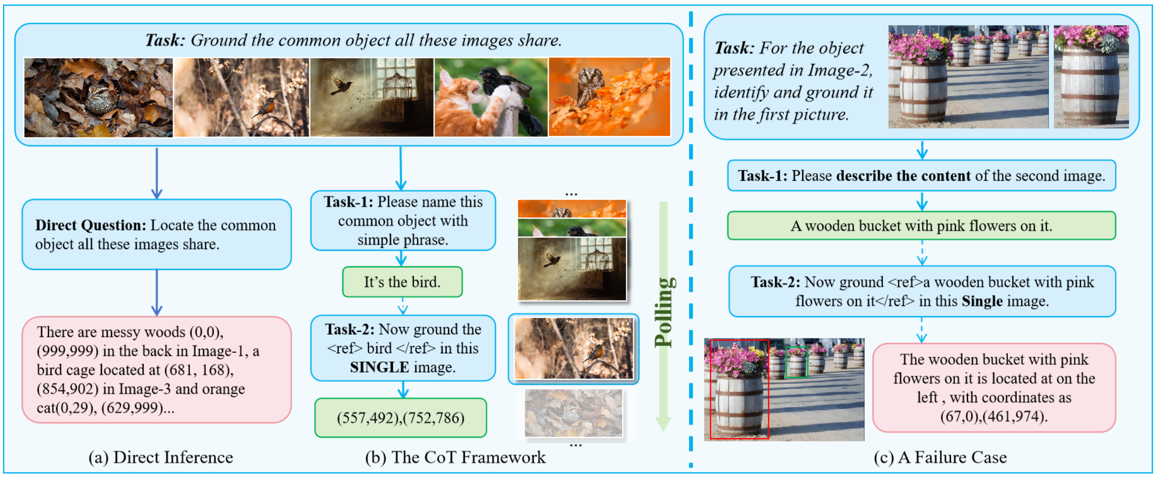


Figure 3: Illustration of the CoT framework and its failure case. Different from (a) direct inference, the (b) CoT method decomposes the task into two subtasks, solving each task deploying the model's existing capabilities. A failure case of CoT is shown in (c) where the model struggles at handling abstract visual information. Green and red background colors indicate correct and incorrect answers, respectively.

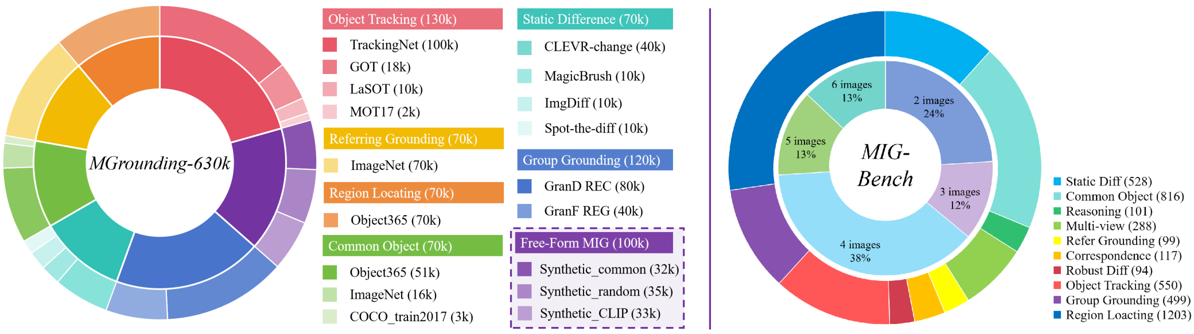


Figure 4: Statistics of the MGrounding-630k dataset and MIG-Bench.

tomatically 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 C.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 D.

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),

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											
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
Minicpm2.6	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	**59.66**	**43.62**	**83.33**	**65.82**	**56.94**	**68.50**	**68.69**	**68.74**	**55.45**	**34.19**	**60.49**

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.

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

To thoroughly assess the multi-image grounding abilities of current MLLMs, we have meticulously curated MIG-Bench. This benchmark consists of 5.9k images and 4.3k testing instances, covering 10 tasks. The distribution of these tasks is illustrated in Figure 4. The benchmark tasks are divided into two categories: spontaneous and referential multi-image grounding. Spontaneous grounding tasks require the model to recognize and ground differences or common objects across images, with a total of 1.4k testing instances. Referential grounding tasks, on the other hand, require the model to utilize different forms of reference queries (i.e., textual, visual, or multimodal) to locate the target objects, comprising 6 distinctive tasks and 2.9k testing instances. The details of these tasks are provided in Figure 2 and Appendix A.

To ensure diversity, the images are sourced from a variety of sources, existing datasets, web images and manually captured photos. For existing datasets, we use examples that exhibits significant movement from GOT-10k_val (Huang et al., 2019) for the Object Tracking task, and manually modify the images from Object365 (Shao et al., 2019) for Common Object task. For Multi-view Grounding, we collect 288 examples spanning both indoor and outdoor scenes from Ego4D (Grauman et al., 2022). The Static Difference task is sourced from the MagicBrush_dev set (Zhang et al., 2024). We also manually capture 97 image pairs with view differences in real-world settings and collect, on average, over 100 image pairs from Google Search for tasks related to Visual Referring, Reasoning, and Referring Grounding tasks.

For a quantitative and objective evaluation of different MLLMs, we ensure that each testing instance contains only one clear target region, eliminating any potential ambiguity. Our MIG-Bench offers a comprehensive evaluation across various real-world scenarios and domains. We believe this benchmark will provide valuable insights into the challenges of MIG and inspire further research in related areas.

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

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
Minicpm2.6	42.65	51.45	*71.09*	69.10	50.19	56.90
Qwen2-VL-7B	42.04	**52.35**	68.06	**70.97**	*54.36*	*57.56*
Migician	**54.27**	*51.50*	69.57	*70.59*	**55.76**	**60.34**

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.

V* Bench	Attribute	Spatial	Overall
Human Level	98.26	100.00	98.95
Random Guess	26.73	50.00	35.99
Tool-using Pipeline			
MM-React	34.78	51.31	41.36
Visprog	31.30	56.57	41.36
SEAL	*74.78*	**76.31**	**75.39**
End-to-end MLLMs			
InternVL2-8B	29.56	56.57	43.07
Gemini Pro	40.86	59.21	48.16
LLaVA-1.5	43.47	56.57	48.68
Minicpm2.6	40.86	64.47	52.67
GPT-4V	51.30	60.53	54.97
Migician$_{zero_shot}$	60.87	61.84	61.36
Migician$_{slice}$	**81.74**	63.16	*72.45*

Table 3: On V* Bench, Migician generalizes well to the hyper-resolution single image in a zero-shot manner.

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.

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 21.61% compared to the second-best model, Qwen2-VL-72B (38.88%), despite having significantly fewer parameters. Note that there is a substantial gap between 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 preliminary grounding capabilities, such as mPLUG-Owl3, InternVL2 series, and Qwen2-VL series, their inherent localization ability provides an implicit 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.

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 9.77%, achieving SOTA performance on MMIU and shows a 1.40% improvement on the large-scale MIBench. We attribute this to the training on a mixture of multi-image understanding and grounding data, which indicates that our proposed MGrounding-630k can enhance general multi-image comprehension.

Setting	Multi-image General Benchmarks					Multi-image Grounding
	MuirBench	BLINK	MIBench	Mantis	MMIU	MIG
Base	42.04	52.35	68.06	70.97	54.36	28.41
Full data	54.27	51.50	69.52	70.51	55.76	53.21
w/o grounding	$44.54_{(-9.73)}$	$51.32_{(-0.18)}$	$71.68_{(+2.16)}$	$67.74_{(-2.77)}$	$52.12_{(-3.64)}$	$22.43_{(-30.78)}$
w/o general	$53.62_{(-0.65)}$	$49.25_{(-2.25)}$	$65.22_{(-4.30)}$	$64.52_{(-6.99)}$	$48.61_{(-7.15)}$	$52.37_{(-0.84)}$

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

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 Visual Search in High-Resolution Images

Finding visual details in high-resolution images is a challenging task, and many recent works have explored this area (Wu and Xie, 2024). Typically, these tasks involve images with very high resolution, where the relevant visual information is often quite small, posing significant challenges to the model's grounding ability. In this section, we analyze and demonstrate that the multi-image grounding capability of Migician can be leveraged to efficiently address this task. Specifically, we slice a single high-resolution image into multiple sub-images and directly transform the problem into a multi-image grounding task. 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 question.

We test this approach on the V*Bench (Wu and Xie, 2024) and list the results in Table 3. In the table, we refer to the method that directly asks Migician to answer the question based on the original image as Migician$_{zero_shot}$, while Migician$_{slice}$ denotes the method that transforms this task into a MIG task as mentioned before. The results remarkably demonstrate the effectiveness of using the MIG approach for high-resolution image visual search. Notably, on the Attribute task, Migician even surpasses the specialized visual searching system SEAL (Wu and Xie, 2024).

7.3 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 the multi-image grounding and 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.

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, there still remain several limitations. First, due to the computational budget, we haven't verified the effectiveness of our training methods on larger 70B scale models. Secondly, in spite of intensive grounding training, our model is still confronted with inaccurate grounding issue, especially in complicated or messy scenarios. 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, 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.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. 2023. One for all: Towards training one graph model for all classification tasks. *arXiv preprint arXiv:2310.00149*.

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

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.

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. Mc-bench: 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 the spontaneous grounding through three distinct tasks below, which aim at assessing model's ability to autonomously discover insidious connections across various images and accurately recognize then locate the target.

Spot the Difference Given two similar images with a single subtle difference, the model is instructed to recognize and ground this difference in the second image, requiring keen perceptual skills.

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

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.

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 a target object within its corresponding specific image. Given a set of images and one textual query, the model must identify the correct image then accurately ground the target object within it.

Visual Reference Query These tasks examine model's ability to effectively utilize visual reference information and incorporate it into the searching 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 locate the referenced object 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 frame, and the model must follow it throughout the sequence.
(4) **Multi-view Grounding.** Here, the model must locate the same target object across multiple images taken from distinct viewpoints.

Visual+Textual Reference Query These tasks combine information from both modalities to assess cross-modal reasoning abilities.
(1) **Correspondence.** The model must ground semantically or functionally similar regions within a 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 Figure2.
Our comprehensive benchmark 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 Single-Image CoT Failure Patterns

As shown in Figure 5, the four representative failure patterns are (a) special multi-image format, (b) abstract visual information, (c) CoT error propagation, (d) step-2 inference error.

When the multiple images are formatted in a special pattern, where our target object is missing in the target image, like, the information in this image is insufficient to perform grounding.

The abstract visual information refers to situations where the intricate visual cannot be adequately converted in textual description to perform accurate grounding. In Figure 5, the simple description "a close-up of a woman's face" cannot distinguish which face the target is in Image-1.

Each reasoning step of CoT could be incorrect, which could potentially leads to the error propagation issue (Yao et al., 2022). In Figure 5, the conclusion draw from the first step is incorrect, which directly leads to the mistake in the second step.

The last failure pattern refers to cases where the erroneous reasoning step appears at the second step, failing to accurately ground or follow instruction.

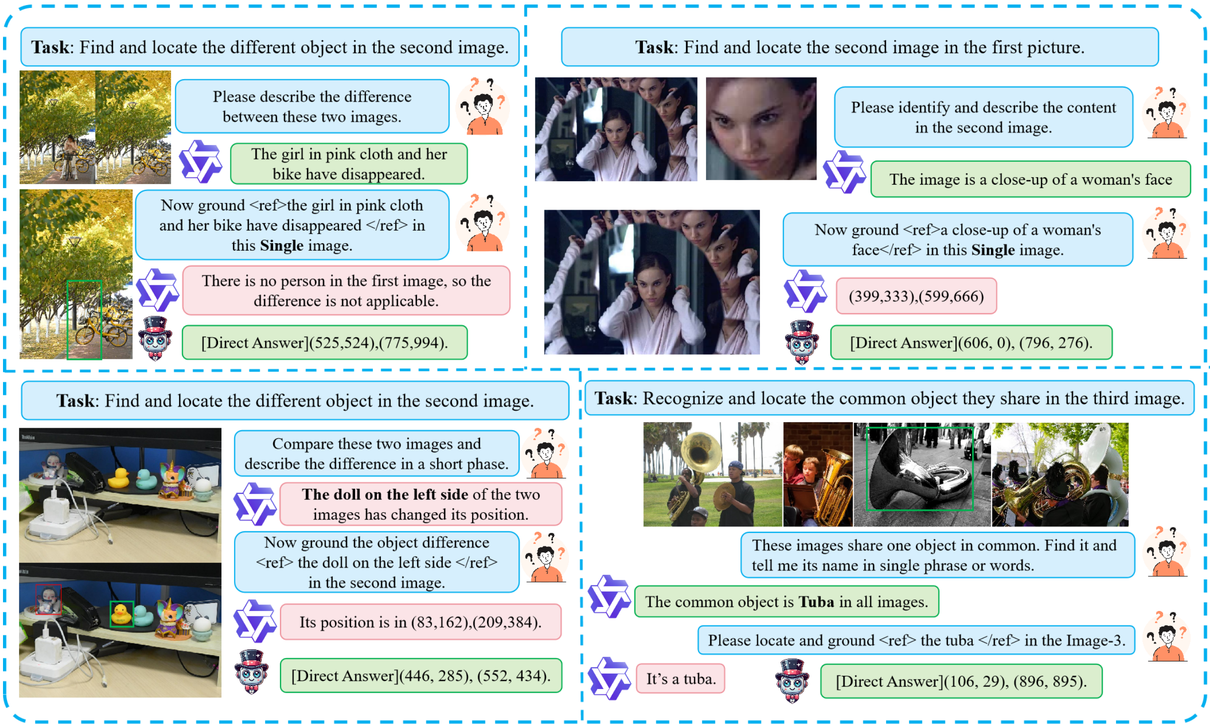


Figure 5: Above are the four representative failure patterns of the single-image CoT. From left to right, top to bottom, they are (a) special multi-image format, (b) abstract visual information, (c) CoT error propagation, (d) step-2 inference error.

C MGrounding-630k Data Curation Details

C.1 Transforming Existing Data

Static Diff Describing the differences of the two nearly same pictures is a well discussed topic, yet they focus on the coarse-grained semantic feature, failing to precisely recognize the part of differences. After comprehensive survey on this area, we have collected high-quality and fully labeled image difference data from 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).

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

Common Object Grounding Recognizing and grounding the main common object in multiple images is an interesting yet non-trivial task for models, which firstly requires them to simultaneous look at multiple images, disentangle the common object, then finally grounding the target object in every single image. We take diverse data sources from ImageNet (Deng et al., 2009), COCO (Lin et al., 2014) and Object365, where the annotations are abundant and rich. Through their annotations, we group the images that share the same object together with a threshold of the proportion it takes for the whole image to filter out too tiny objects. We empirically find such threshold effective at eliminating ambiguity where there could be multiple common object candidates, which results in clear and definite training examples. We further reduce ambiguity by skipping inappropriate classes where there are always multiple possible candiate such as couch, dinning table, keyboard and etc.

Object Tracking The original object tracking emphasize more on tracking the target object in a video sequence, which resembles with multi-image sequences. As a well discussed topic, we select the large scale TrackingNet (Muller et al., 2018), LaSOT (Fan et al., 2019), GOT-10K (Huang et al., 2019) and MOT-2017 (Milan, 2016) datasets as our data sources. During the dataset construction process, we have simplified the original long sequence to 4-6 image per training example, maintaining its original core feature while keeping efficiency. For each image, we extract the key frame with an ap-

Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
KOSMOS-2 (Peng et al., 2023)	52.32	57.42	47.26	45.48	50.73	42.24	60.57	61.65
VisionLLM v2 (Wu et al., 2024)	79.20	82.30	77.00	68.90	75.80	61.80	73.30	74.80
OFA-L (Liu et al., 2023)	80.00	83.70	76.40	68.30	76.00	61.80	67.60	67.60
Shikra (Chen et al., 2023b)	87.00	90.60	80.20	81.60	87.40	72.10	82.30	82.20
InternVL2-8B (Cai et al., 2024)	87.10	91.10	80.70	79.80	87.90	71.40	82.70	82.70
GroundingGPT (Li et al., 2024d)	88.02	91.55	82.47	81.61	87.18	73.18	81.67	81.99
Griffon v2 (Zhan et al., 2024)	**89.60**	**91.80**	**86.50**	81.90	85.50	**76.20**	**85.00**	**86.00**
Migician	88.48	91.51	82.85	**82.94**	**88.23**	75.50	84.15	83.67

Table 6: Performance of Models on single image grounding benchmark.

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 7: 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.

propriate interval to ensure the obvious movement of the key target. We also involve small proportion of the ordering judge for continuous images to enhance the model's temporal understanding ability.

Referring Grounding This part of training data is designed to imitate finding the object of the source image in the target image. We mainly utilize the ImageNet-2012 subset to construct image pairs, with the first one taking large proportion of the whole image, and the second one containing smaller target object that may take efforts to spot. In total, the refering grounding dataset covers a wide range of objects that is beneficial for the model to generalize.

Group Grounding Conventional visual grounding is mostly limited to single image context, while iin real world scenario, we often need to recognize the targte object from a messy piles of pictures. Group Grounding, which locates the target among a group of different images, is the exact task to fill in such gap and enrich the versatility of traditional grounding. When constructing this part of data, we take advantage of the single image GranD rec and reg conversation data (Rasheed et al., 2024) with the quantity of 3M. With further filtering and combining 3-5 images per group, we finally obtained a collection of 12w high-quality training data for stage-1 training(grounding injection training), which is effectively at enhancing the co-reference and image-level locating ability of models.

Region Locating Region locating refers splitting several pieces of semantic-rich region from the source image and then recognize the exact location of this region image in the source picture. To guarantee extracting meaningful regions, we utilize the Object365 dataset and extract the bounding box area as the regions. To further improve quality, we set a series of filtering mechanism:(1) content richness: we select the images that have more than 10 bounding box annotations to avoid too plain or simple cases. (2) aspect ratio: we keep the aspect ratio between 0.5-2 to avoid the excessively thin bounding box areas that models may fail to effectively tackle. (3) size: we keep the region ratio of the whole image between 0.2-0.49 and an absolute pixels above 2000 to avoid excessively tiny and obscure region images. Noticeably, due to our meticulously crafted mechanism and the feature of this task, the resulting training data encompasses the cases of person recognition in the source image, analogous objects distinguishing and tiny details recognition that are non-trivial even for human.

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
Migician	Polling	81.99	44.44	61.09	59.65
Migician	All	72.43	43.06	58.55	34.91

Table 8: Comparison of different answering forms. For random guess, we set the default answer as *(0,0),(999,999)*.

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 9: 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 10: Training data proportion for two stages.

Algorithm 1 CLIP Adaptive Similarity Selection

Require: Images $\mathbf{I}$, adaptive selection range k, $thres \in (0, 1)$
Ensure: Final Image Set $\mathbf{F}$

Initialize $\mathbf{F} \leftarrow \emptyset$
Extract $\mathbf{F}_I \leftarrow$ Features of $\mathbf{I}$
while $\mathbf{F}_I$ is not empty **do**
  Randomly select $thres \sim \text{Uniform}(0.1, 1)$
  for each $f_i \in \mathbf{F}_I$ **do**
    $s_{ij} = \text{similarity}(f_i, f_j), \forall f_j \in \mathbf{F}_I, j \neq i$
  end for
  $\textbf{Sort\_S}_i = \text{Sort}(s_{ij})[1:]$
  $k \leftarrow \lfloor thres \times (\text{len}(\textbf{Sort\_S}_i)) \rfloor$
  $\textbf{Candidates} \leftarrow \textbf{Sort\_S}_i[:k]$
  Randomly select $r \sim \text{Uniform}(3, 5)$
  $\textbf{Selected} \leftarrow \text{Sample}(\textbf{Candidates}, r)$
  Append f_i and **Selected** to $\mathbf{F}$
  Remove f_i and **Selected** from $\mathbf{F}_I$
end while
return $\mathbf{F}$

C.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.

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).

D Details of Two-Stage Training

This section outlines the data proportions and their respective sources for the two training stages, as summarized in Table 10.

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.

E Evaluation Implementation

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.

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 8, Migician still demonstrates great robustness to the variation of evaluation format.

The performance of Migician is presented in Table 6. Although mainly targeted at mutli-image grounding, Migician still maintains well on conventional single-image grounding task.

70B Scale Models The performances of three competitive 70B scale models are illustrated in Table 9 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.

F 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 7. 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.

G Case Study

We provide detailed cases comprehensively reflecting the free-form MIG ability of Migician in Figure 6, 7, as well as our instruction tuning data details examples in Figure 8.

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.

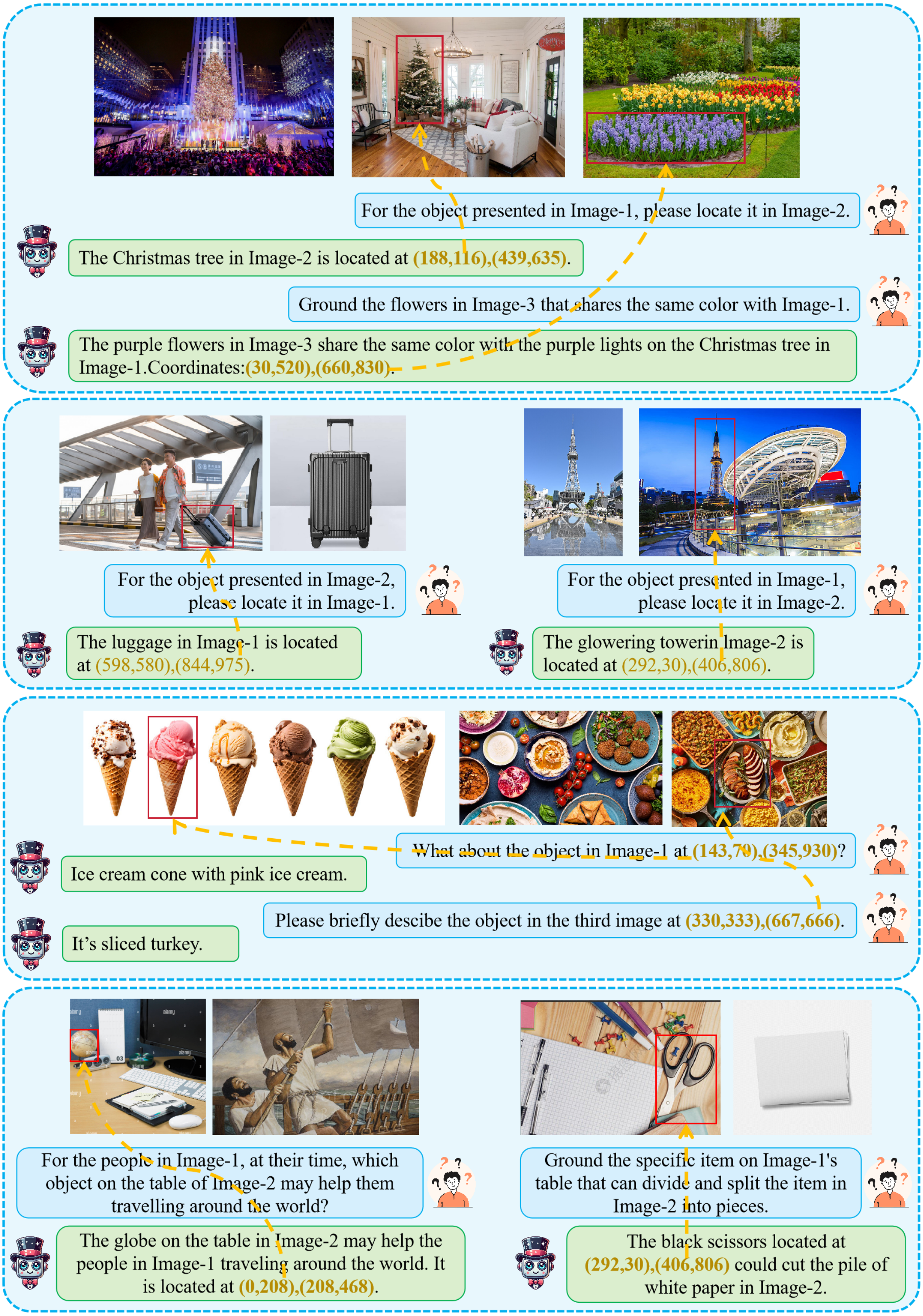


Figure 6: Example cases of the free-form multi-image grounding ability of Migician.

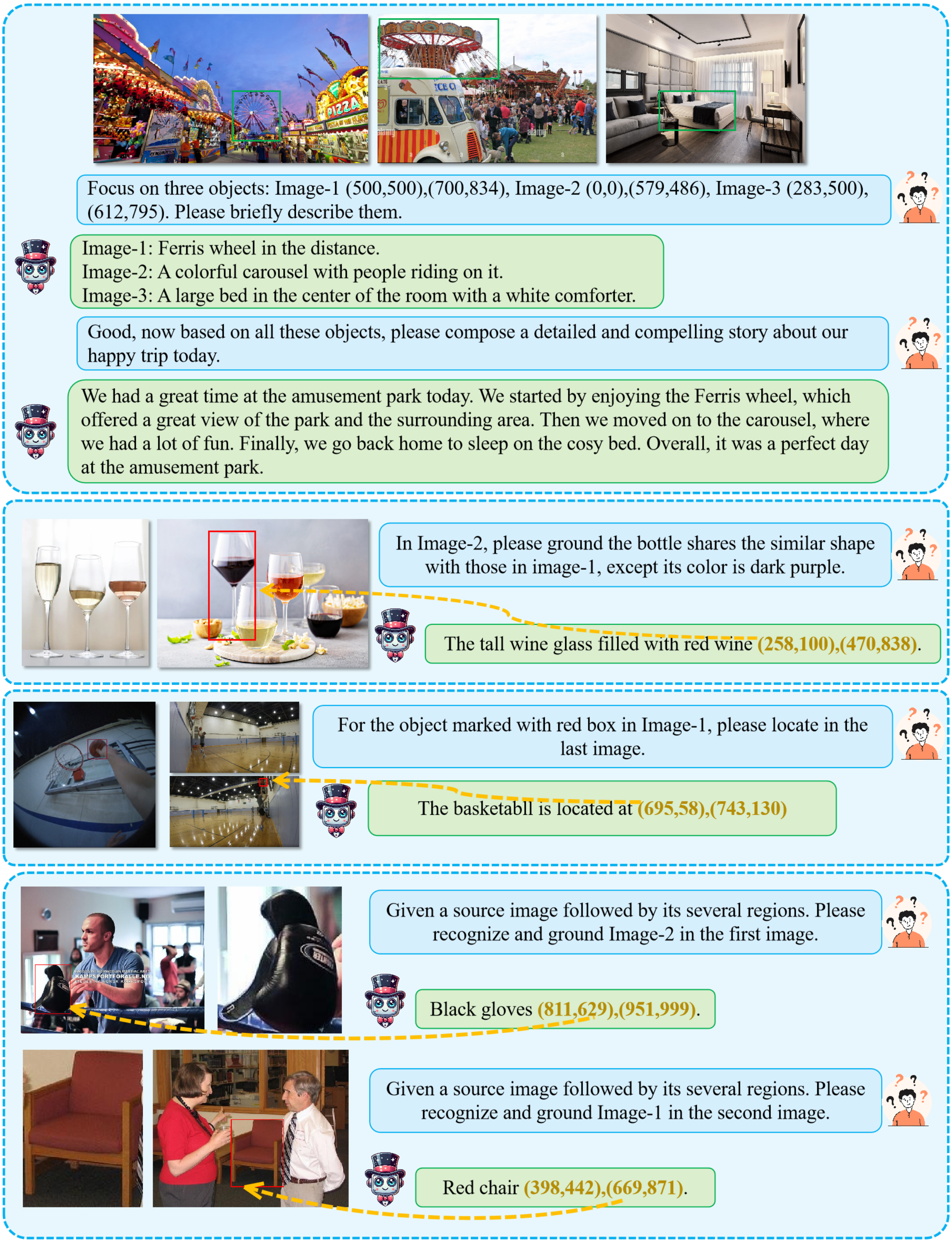
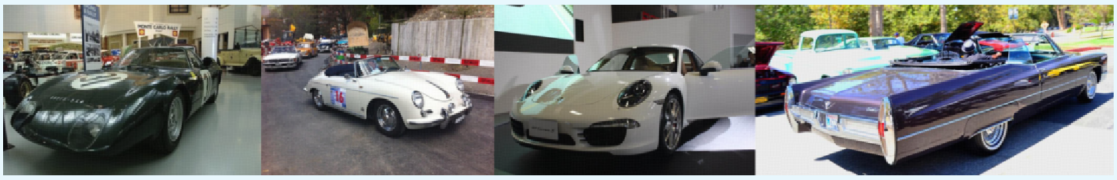


Figure 7: Example cases of the free-form multi-image grounding ability of Migician.



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