MMIFEvol: Towards Evolutionary Multimodal Instruction Following

Haoyu Wang[1], Sihang Jiang[1]*, Xiangru Zhu[1], Yuyan Chen[1,3], Xiaojun Meng[2], Jiansheng Wei[2], Yitong Wang[1], Yanghua Xiao[1]*

[1]Shanghai Key Laboratory of Data Science, College of Computer Science and Artificial Intelligence, Fudan University,
[2]Huawei Large Model Data Technology Lab,
[3]Cornell University
{wanghy24@m., jiangsihang@, xrzhu19@, yitongw@, shawyh@}fudan.edu.cn,
{xiaojun.meng, weijiansheng}@huawei.com, yolandachen0313@gmail.com

Abstract

Multimodal Instruction Following serves as a fundamental capability of multimodal language models, involving accurate comprehension and execution of user-provided instructions. However, existing multimodal instruction-following datasets and benchmarks face the shortcomings outlined below: (a) *Lack of Difficulty Stratification*, they collect diverse instruction categories but neglect the stratification of difficulty levels across these categories, which leads to overlap, bias, and low interpretability. (b) *Lack of Fine-Grained Metrics*, they conflate the model's ability to "solve tasks" and "follow constraints" into a single metric, which fails to accurately reflect its instruction-following capability. (c) *Lack of Multi-Task Instructions*, they overlook the fact that real-world user instructions often consist of multiple combined tasks. This paper proposes **MMIFEvol**, a framework for multimodal instruction evolving and benchmarking. First, we define the essential components of a carefully curated multimodal instruction set and establish corresponding difficulty levels, based on which we synthesize diverse instruction data. Next, we decouple the evaluation criteria for the instruction following into three different metrics to construct a high-quality benchmark and assess existing models. Experimental results demonstrate that current models still struggle with following complex instructions, while fine-tuning using MMIFEvol data effectively improves models' responsiveness to multimodal instructions.

Introduction

Research on Multimodal Large Language Models (**MLLMs**) (Wu et al. 2023; Bai et al. 2024) has achieved remarkable performance in image understanding (Li et al. 2024a), image-based text generation (Dong et al. 2023), and visual reasoning (Xu et al. 2024) by constructing multimodal instructions and performing instruction tuning (Liu et al. 2023). Instruction Following (**IF**) refers to a model's ability to understand visual input and solve specific tasks as required (Su et al. 2023; Ren et al. 2025), which is crucial for advanced MLLMs, from chatbots to practical productivity tools for real-world applications (Lei et al. 2024). For example, in an OCR scenario (Chen et al. 2025), when a user instructs the MLLM to `parse the text in the image and return it in JSON format`, a precise IF requires the MLLM not only to accurately recognize the text but also to correctly return formatted content (e.g. `{ 'content': 'text' }`) rather than free-form text, ensuring seamless integration with the user's downstream applications. Consequently, with the rapid advancement of MLLM capabilities, evaluating and improving models' multimodal instruction-following

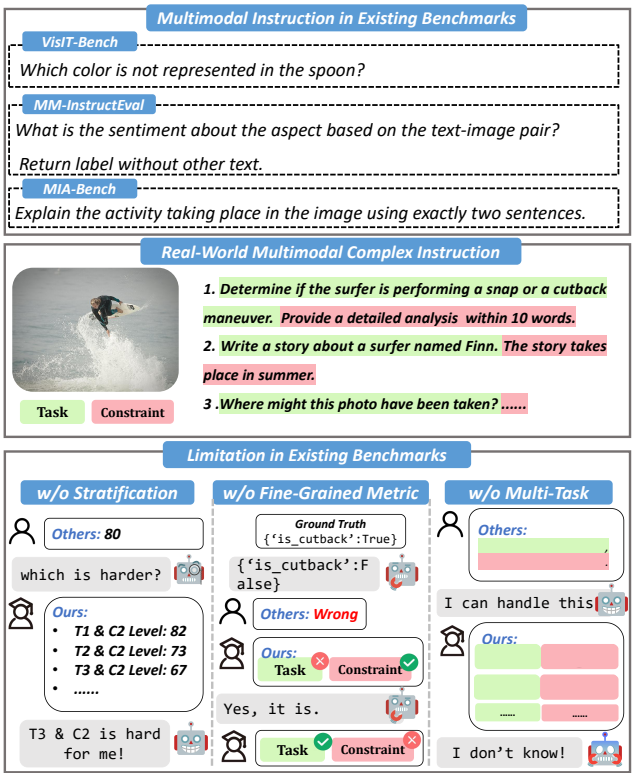

Figure 1: Multimodal instructions consist of ***input***, ***task***, and ***constraint***. However, existing benchmarks fail to systematically categorize the difficulty of tasks and constraints, also do not distinguish between "solving tasks" and "following constraints", and overlook the complex instruction with multiple tasks and constraints.

*Corresponding Author

abilities has become a critical research imperative.

Existing benchmarks for multimodal instruction following remains in its infancy, which can be categorized into two main types: (a) Task-Centric methods (Bitton et al. 2023; Yang et al. 2025), which select diverse visual tasks from multimodal pre-training datasets and rewrite them using advanced models such as GPT-4o (Hurst et al. 2024); and (b) Constraint-Centric methods (Qian et al. 2024; Ding et al. 2025), which predefine multiple types of constraints and rewrite instructions by matching and integrating task-relevant constraints. However, these multimodal instruction-following evaluations exhibit limitations, as illustrated in Figure 1:

- **Lack of Difficulty Stratification**. Existing benchmarks collect several categories of multimodal tasks or constraints for dataset construction, but fail to account for the hierarchical difficulty levels among these categories during evaluation. This oversight leads to issues of overlap, bias, and diminished interpretability. In contrast, evaluations with well-defined difficulty stratification can more precisely identify a model's capability boundaries, thereby enabling targeted supplementation of training data to enhance specific model competencies in a focused manner.
- **Lack of Fine-Grained Metrics**. Existing benchmarks tend to conflate the model's ability to "solve tasks" and "follow constraints" into a single metric, which fails to accurately reflect its instruction-following capability. In practice, MLLMs may be capable of solving specific visual tasks but struggle with adhering to associated constraints.
- **Lack of Multi-Task Instructions**. Existing benchmarks typically construct instructions with single-task, yet have not explored real-world scenarios involving multiple tasks and multiple constraints. In practice, models may perform well on individual tasks but struggle with composite instructions involving multiple tasks.

To construct high-quality multimodal instruction following benchmark and accurately evaluate the IF abilities of existing MLLMs, we propose **M**ulti-**M**odal **I**nstruction **F**ollowing **Evol**ution, a framework for multimodal instruction evolving and multimodal IF evaluation, as depicted in Figure 2. Specifically, we decouple multimodal instructions into three core components: *input*, *task*, and *constraint*. For input, we apply clarity-based and content-based filtering to image sources, retaining images rich in semantic information. For task, building on principles from cognitive science, we establish the first difficulty stratification of multimodal capabilities and develop a diverse visual task pool based on this taxonomy. For constraint, extending concepts from textual instruction following, we construct a structured constraint hierarchy and corresponding constraint pool, incorporating cross-modal constraints specific to multimodal settings. Utilizing well-defined difficulty stratification, we leverage powerful MLLMs to automatically generate tasks and constraints based on image content, thereby enhancing the interpretability of instruction generation and evolution. Furthermore, we integrate complex instructions involving multiple tasks and constraints, while employing automated quality control to filter out low-difficulty synthetic instructions. Finally, we formulate three fine-grained metrics: *accuracy* ($\mathcal{A}$), *following* ($\mathcal{F}$), and *preference* ($\mathcal{P}$), to holistically evaluate multimodal IF performance in both open-source and closed-source models. Experimental results demonstrate that:

a) **MLLMs exhibit notable deficiencies in handling tasks and constraints at specific difficulty levels**. For the same task, the model may generate hallucinations due to changes in constraints; some models also demonstrate difficulties in adhering to certain types of constraints. This substantiates the necessity of hierarchical difficulty-based evaluation.

b) **The models' ability in task-solving is not equivalent to their ability in constraint-following.**. Previous evaluations have not sufficiently decoupled these two aspects, resulting in a potentially biased or incomplete assessment of model capabilities.

c) **The models exhibit poor performance when faced with complex instructions involving multi-tasks and multi-constraints**. As the number of tasks and constraints increases, the model performance on complex instructions decreases.

Our contributions can be summarized as follows:

- We propose MMIFEvol, a framework for multimodal instruction evolving and evaluation of multimodal instruction following. Based on cognitive science, we come up with the first difficulty stratification of multimodal capabilities and evolve diverse and challenging multimodal instructions accordingly.
- We establish three fine-grained evaluation metrics for multimodal instruction following, construct a high-quality benchmark, and assess mainstream MLLMs. Experimental results reveal an inconsistency between models' "task-solving" and "constraint-following" abilities, a critical aspect overlooked in prior evaluations, and existing models still struggle with following complex instructions.
- We further fine-tune models using instructions synthesized by MMIFEvol, observing significant performance improvements on both our benchmark and other multimodal instruction-following benchmarks. This validates the effectiveness of our instruction evolution and evaluation methodology.

Related Work

Multimodal Instruction Generation

LLaVA (Liu et al. 2023) enhances the visual understanding capabilities of multimodal language models by utilizing LLaVA-Instruct, a dataset annotated by advanced LLMs. ALLaVA (Chen et al. 2024) enhances diversity by manually crafting and rewriting instruction data. MMInstruct (Liu et al. 2024) leverages advanced MLLMs to automatically generate instruction data, thereby improving the efficiency of data synthesis. MMEvol (Luo et al.

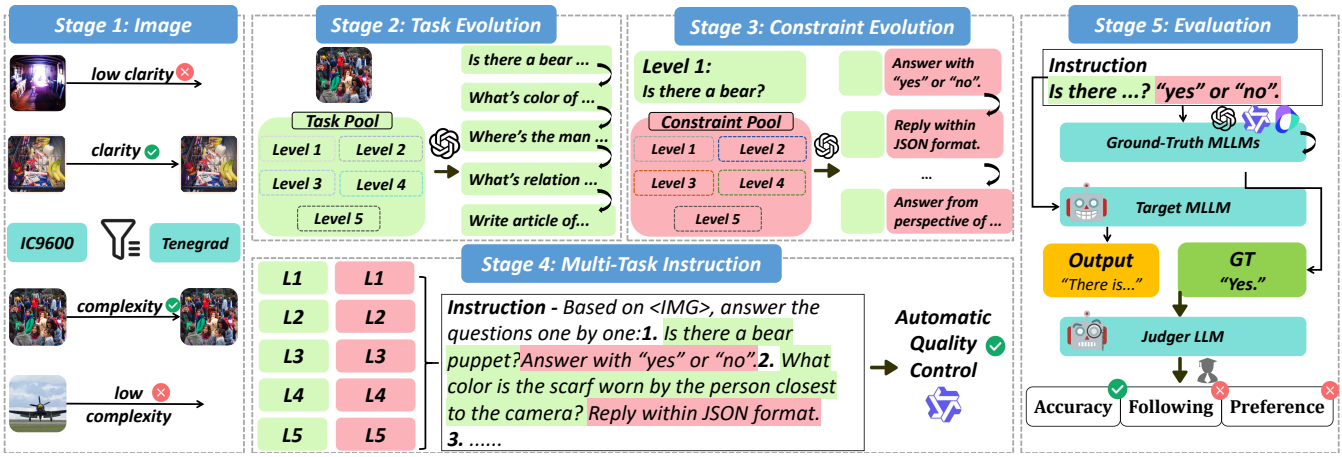

Figure 2: The framework of our benchmark design contains five stages for evolving and evaluating multimodal instructions, an evaluation dataset covering five levels of task and five levels of constraint, and three fine-grained evaluation criteria.

Benchmark	Format	Taxonomy	Complexity	Metric Granularity
VisIT-Bench	VQA	None	1 Task	Task-Accuracy
MM-InstructEval	Multiple Chioce	6 Task Categories	1 Task	Task-Accuracy
MIA-Bench	Freeform	8 Constraint Categories	1 Task + n Constraints	Constraint-Following
MMIF-Eval	Freeform	6 Constraint Categories	1 Task + n Constraints	Constraint-Following
MMIFEvol	Freeform	25 Task $\times$ Constraint Levels	n Tasks + n Constraints	$\mathcal{A}$, $\mathcal{F}$, $\mathcal{P}$

Table 1: Comparison of existing multimodal instruction following benchmarks.

2024) evolves the LLaVA-Pretrain-595K dataset along three dimensions—fine-grained perception, cognitive reasoning, and interaction—thus increasing the difficulty of the instructions. Oasis (Zhang et al. 2025) enables automatic and diverse multi-modal instruction data generation from single image, significantly improving MLLM performance without requiring manual annotation.

Multimodal Instruction Following Benchmark

The evaluation of instruction following in multimodal settings has emerged as a recent research focus. As an early research, VisIT-Bench (Bitton et al. 2023) constructed 592 challenging visual instructions, with responses generated by GPT-4 based on image captions and subsequently verified by human experts. MM-InstructEval (Yang et al. 2025) perform extensive zero-shot evaluations that include 6 distinct tasks using 10 different instruction templates. MIA-Bench (Qian et al. 2024) starts with eight predefined instruction categories, rewrites the seed instructions, and uses a fine-grained model-based evaluation framework to assess the degree to which various constraints within the instructions are followed. MM-IFEval (Ding et al. 2025) designs both Compose-Level and Perception-Level questions to construct a diverse set of visual instructions. Compared to existing benchmarks, our evaluation demonstrates advancements in the taxonomy of instruction synthesis, the granularity of evaluation metrics, and the complexity of instructions, as shown in Table 1.

Method

We propose MMIFEvol, a framework designed to evolve instructions of varying difficulty levels from a single image, as illustrated in Figure 2. The entire framework consists of five stages: (1) Image Selection, which aims to obtain high-quality images rich in semantic information; (2) Task Evolution, where visual tasks with clear difficulty stratifications are generated based on the image content; (3) Constraint Evolution, which adds constraints to the image-task pairs to modulate the difficulty; (4) Multi-Task Instruction Integration, in which multiple task-constraint pairs are rewritten into one complex instruction; and (5) Evaluation, in which scoring LLMs are employed to assess the quality and precision of model responses across three fine-grained metrics: accuracy, following, and preference.

Image Selection

Images in complex multimodal instructions typically contain rich visual information. Therefore, considering the subsequent instruction evolution process, we filter images from two perspectives: clarity and complexity, aiming to obtain a high-quality image set. We prioritize real-world images (e.g. Microsoft COCO (Lin et al. 2014)) due to their rich semantic entities, logical relationships, and cultural contexts. This enables models to more freely generate diverse tasks and provides ample space for creation in subsequent instruction evolution processes.

Specifically, given the original image set $\hat{I}$, the category

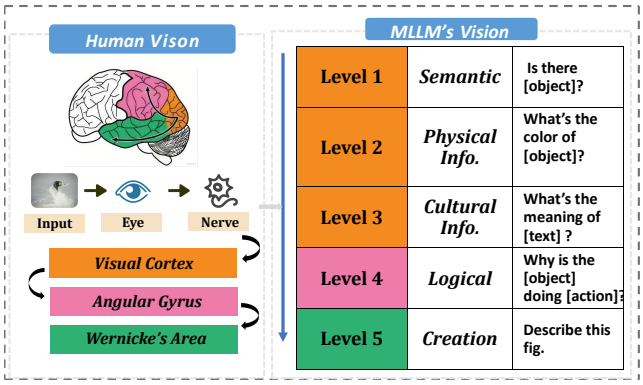

Figure 3: Criteria for Task Evolution. Analogous to the human visual processing pathway (visual cortex – angular gyrus – Wernicke's area), we define five levels of task with progressively increasing difficulty and construct corresponding task pools. By prompting the model to complete tasks based on image content and task templates, we synthesize preliminary instruction data with clearly defined levels of difficulty.

set C, and the target category c, for each image $i \in \hat{I}_c$, we compute its **sharp** score and **complexity** score based on the Tenengrad (Tenenbaum 1971) sharpness and the pre-trained scoring model IC9600 (Feng et al. 2022) :

$$S_s(i) = \frac{1}{N} \sum_{(x,y)} \left(G_x(x,y)^2 + G_y(x,y)^2 \right)$$

$$S_c(i) = f_{IC9600}(i)$$

where $G_x$and G_y are the horizontal and vertical gradients calculated by the Sobel operator on the grayscale image. The effectiveness of these two methods in analyzing the clarity and complexity of images has been widely verified. Subsequently, we calculate the total image quality, then sort and filter the images with a ratio of p:

$$S(i) = \alpha S_s(i) + \beta S_c(i)$$

$$I_c = \{ i \mid S(i) \geq S_p^*, i \in \hat{I}_c \}$$

where S_p^* is the score of the first $\lceil n \cdot p \rceil$ image after sorting all the images in category c from high to low by $S(i)$. Ultimately, by integrating various types of images, an image set $I = \bigcup_{c \in C} I_c$ is obtained for the subsequent evolution of the instruction.

Task Evolution

To evaluate MLLMs' instruction-following capabilities and conduct targeted capability diagnostics, it is necessary to define and categorize the tasks involved in multimodal instruction following. Drawing on the structure of the human brain (Grill-Spector and Malach 2004; Raichle and Mintun 2006) and the framework of human cognitive abilities (Forehand 2010) , we propose a hierarchical five-level categorization of multimodal capabilities, with corresponding tasks designed for each level, as illustrated in Figure 3. This hierarchy progresses from the matching of basic semantic entities and the recognition of physical attributes to higher-level semantic understanding, logical reasoning, and creativity. Guided by this classification system, we have designed a comprehensive task pool for each level. This approach reduces intercategory overlap and imbalance in data distribution and enhances the interpretability of evaluation results.

Based on this criterion, we use the closed-source GPT-4o model (Hurst et al. 2024) to generate tasks according to the content of the image. Specifically, for an image $i \in I$ and task difficulty $d \in \{1, 2, 3, 4, 5\}$, we collect task templates from the level-d task pool and prompt the model to select the most suitable template for instantiation based on image content. The diverse task templates defined by experts prevent the model from generating homogeneous instructions, while image-based instantiation ensures the uniqueness of each instruction. Furthermore, to increase the difficulty of the tasks and ensure the robustness of the evaluation, we prompt GPT, with probability p, to generate questions that include semantic entities not present in the image, in order to observe possible hallucinations exhibited by the model to be evaluated. In conclusion, for each image, we construct a set of tasks with explicit difficulty levels, denoted $T = \{t[i, d]\}_{i \in I,\, d \in \{1,2,3,4,5\}}$.

Difficulty	Category	Constraint Case
Level 1	Content	*Answer [**"Yes"**] or [**"No"**].*
Level 2	Format	*Reply within [**json**] format.*
Level 3	Count	*Reply within **3** words or less.*
Level 4	Situation	*Focus on [**0.55,0.25,0.85,0.35**].*
Level 5	Style	*Answer in a [**humorous**] tone.*

Table 2: Criteria for Constraint Evolution

Constraint Evolution

Similarly to task evolution, in order to evaluate the ability of visual models to follow various types of constraints, we generate constraints with explicit difficulty levels for each task. The categorization of constraints in language model instructions has been extensively studied in the field of natural language processing. Therefore, we adopt the definition of constraint categories from FollowBench (Jiang et al. 2023) - whose validity has been widely verified and design a corresponding constraint pool based on this definition, as shown in Table 2. The difficulty of following constraints increases progressively, ranging from content, format, and quantity constraints, which are relatively easy to execute and verify, to more abstract and higher-order constraints, such as state and style. Considering the unique characteristics of multimodal instructions compared to text-only instructions, we incorporate several multimodal-specific constraints into the pool, such as requiring the model to focus on specific regions (grounding boxes) within the image.

We employ GPT-4o to perform the instruction evolving for this stage. Specifically, for the set of tasks T, we prompt the model to rewrite the original tasks according to the definitions and the constraint pool: $I_s[i, d_t, d_c] =$

$LLM(t[i, d_t], \mathtt{P})$, where P is the prompt template, d_t and d_c respectively represent the difficulty levels corresponding to task and constraint. As a result, we generate the corresponding instruction data for each image, each difficulty level of the task, and each difficulty level of the constraint.

Multi-Task Instruction Integration

In real-world scenarios, user instructions are often complex and can simultaneously involve multiple tasks and constraints, a characteristic that has been overlooked in previous evaluations. Compared to simple instructions containing only a single task and constraint, complex instructions impose higher demands on a model's contextual understanding, memory, and execution capabilities. Therefore, we extend the constructed single-task instruction set I_s to a complex multi-task instruction set I_m, in which multiple tasks and constraints are combined. Specifically, for each image i and its corresponding atomic instruction $I_s[i, d_t, d_c]$ (where each atomic instruction consists of a task and a constraint), the complex instruction can be formulated as follows:

$$I_m[i] = \bigcup_{d \in \{1,2,3,4,5\}} I_s[i, d, d]$$

Furthermore, to automatically ensure the quality of synthetic instructions, we evaluated the Visual Instruction Following Difficulty (VIFD) (Li et al. 2023) of the instructions based on Qwen2.5-VL-3B ((Bai et al. 2025)):

$$VIFD(Q) = \exp\left(\frac{1}{N}\sum_{t=1}^{N}\log\frac{P(a_t \mid a_{<t}, I)}{P(a_t \mid a_{<t}, I, Q)}\right) = \frac{\text{Perplexity of Answer given Image \& Query}}{\text{Perplexity of Answer given Image}}$$

and filter out instructions with low VIFD scores to guarantee the difficulty of the final instructions set.

Evaluation System

Existing research on instruction following (Bitton et al. 2023; Qian et al. 2024; Yang et al. 2025) typically focuses solely on the correctness of the model's response, without decoupling "solving task" from "following constraint". Conflating these two aspects leads to limitations in evaluating the instruction-following capabilities of multimodal models. Furthermore, since the ultimate goal of instruction following is to enhance user experience, users' subjective preference of model responses should also be considered as an integral part of assessing instruction-following performance.

Taking these factors into account, we designed three fine-grained metrics for MMIFEvol:

- *Accuracy ($\mathcal{A}$): Does the model correctly answer the question specified in the instruction?* In this metric, we disregard the format, language, style, and other formal aspects of the model-generated content, focusing solely on whether the model correctly addresses the query specified in the instruction. We define four tiers (0, 25, 75, 100) representing the extent to which the model accurately accomplishes the task. Since the correctness of level 5 tasks (creation) is inherently challenging to evaluate, we exclude them from the calculation of this metric.
- *Following ($\mathcal{F}$): Does the model accurately adhere to the constraints and limitations outlined in the instruction?* In this metric, we disregard the correctness of the model's response to the query in the instruction, focusing instead on whether the model strictly adheres to the specified constraints regarding format, quantity, style, etc. We define four levels (0, 25, 75, 100) to represent the degree to which the model follows the constraints.
- *Preference ($\mathcal{P}$): Compared to the Ground Truth, is the model's response better or worse?* The evaluations of $\mathcal{A}$ and $\mathcal{F}$ have objective ground truth, but in actual user interaction, users' subjective feelings are also an important part of the response effect of instructions. Therefore, after two objective metrics, we add a subjective evaluation metric, $\mathcal{P}$, which represents the win rate of the model-generated response (0 or 100). We defined preference from three perspectives: *completeness*, *fluency* and *tone*.

Multi-task instruction metrics are computed as the average of constituent single-task scores. To achieve accurate and fine-grained evaluation, two key elements must be addressed: generating correct ground truth for instructions and correctly calculating metrics.

- For ground truth generation, GPT-4o (Hurst et al. 2024) produces preliminary answers, which are then verified by other leading vision-capable models, Seed1.5-VL (Guo et al. 2025) and Qwen2.5-VL-72B (Bai et al. 2025) - through consistency voting to mitigate potential hallucinations or instruction-following errors.
- For metrics, considering cost, consistency, and robustness in instruction following, we employ CompassJudger (Cao et al. 2024) as a scoring model, a specialized LLM trained for comparative evaluation. Guided by few-shot and chain-of-thought prompting, this judger achieves high consistency with human annotation in evaluating instruction following.

Experiment

Experimental Setup

We evaluated seven main multimodal large language models that have demonstrated outstanding performance on other benchmarks (Team et al. 2023; Grattafiori et al. 2024; Li et al. 2024a,b; Team et al. 2025; Bai et al. 2025; Zhu et al. 2025) . Both inference and fine-tuning were conducted using the MS-Swift (Zhao et al. 2025) framework, all experiments were performed on workstations equipped with 8 $\times$ A800 GPUs.

Results

RQ1 What's the overall performance of existing MLLMs in multimodal instruction following? As shown in Table 4 , Gemini surpasses mainstream open-source models across all three metrics ($\mathcal{A}, \mathcal{F}, \mathcal{P}$), demonstrating its superior visual understanding and instruction-following capabilities. Notably, Gemini exhibits the most significant advantage in Preference, while showing a relatively smaller gap with the top-performing open-source model InternVL3 in

Model		Accuracy ($\mathcal{A}$)				Following ($\mathcal{F}$)					Preference ($\mathcal{P}$)				
		T1	T2	T3	T4	T1	T2	T3	T4	T5	T1	T2	T3	T4	T5
Close-Source MLLMs															
Gemini2.0-Flash	C1	98.2	92.4	97.9	87.5	86.8	77.2	86.5	100.0	100.0	80.6	75.4	87.1	88.2	95.1
	C2	82.4	73.1	67.0	88.8	92.6	75.0	90.0	93.8	97.1	84.1	40.8	30.0	87.5	92.4
	C3	89.2	73.7	70.2	75.0	62.5	56.6	43.3	71.2	93.8	54.2	47.4	23.1	36.5	75.7
	C4	86.2	75.0	72.4	80.8	95.0	82.9	96.1	96.1	100.0	95.0	78.9	72.6	94.7	98.8
	C5	90.0	87.5	80.3	86.9	95.0	95.3	85.5	98.4	96.4	92.6	91.3	79.5	91.5	88.7
Open-Source MLLMs															
LLaVA-NeXT-7B	C1	84.9	93.1	97.0	76.3	97.4	85.3	23.7	87.7	46.1	72.4	56.9	22.4	59.6	12.3
	C2	61.8	55.7	53.1	58.8	98.2	50.9	34.8	96.5	53.1	56.1	12.3	8.8	36.8	19.3
	C3	40.8	44.3	37.7	45.2	91.2	25.3	30.7	82.5	44.3	38.6	10.5	14.0	33.3	7.0
	C4	62.7	39.5	52.2	47.4	90.4	48.2	53.9	75.4	40.8	43.9	21.1	26.3	21.1	8.8
	C5	58.3	73.7	82.5	43.3	78.5	92.1	82.5	68.3	45.1	3.5	29.8	42.1	5.4	5.4
Qwen2.5-VL-7B	C1	96.2	53.4	98.5	93.6	98.5	59.7	95.71	84.1	82.6	82.7	5.3	96.2	57.6	58.3
	C2	72.9	7.0	71.0	68.6	99.4	36.7	93.2	83.1	78.8	59.1	1.5	60.6	53.8	41.7
	C3	69.7	22.2	58.3	65.0	98.1	45.8	86.9	83.9	80.9	58.3	2.3	47.7	54.5	28.0
	C4	77.1	6.6	75.2	80.3	97.7	23.9	92.7	92.4	84.0	66.7	3.0	58.8	74.0	26.0
	C5	91.2	81.9	93.1	83.8	98.3	94.1	94.5	97.3	82.8	26.0	28.2	61.8	18.3	17.6
InternVL3-8B	C1	88.4	70.1	99.2	86.8	90.9	87.7	98.8	88.8	81.5	62.2	28.7	29.1	53.5	77.6
	C2	77.2	66.7	71.5	68.9	86.8	77.9	86.8	62.1	73.8	30.7	35.7	29.5	31.9	55.1
	C3	59.4	51.4	61.8	57.7	81.9	70.7	84.4	73.0	73.0	39.0	31.8	29.5	36.2	50.4
	C4	74.0	52.6	71.3	71.5	79.7	76.3	84.8	79.5	76.4	59.1	33.8	45.7	65.7	66.7
	C5	76.2	65.7	82.3	75.1	80.8	70.8	90.9	80.0	73.4	82.1	45.8	52.9	76.6	53.6

Table 3: The performance of MLLMs in responding to single instructions formed by combinations of tasks (T1–T5) and constraints (C1–C5). **Accuracy** reflects the model's ability to correctly solve the problem posed in the instruction, **Following** indicates the extent to which the model adheres to the constraints specified in the instruction, and **Preference** represents users' subjective evaluation of the model's response. Higher values and more vivid colors in the table represent stronger model performance on instructions of corresponding difficulty levels.

Model	Single-Task Inst.			Multi-Task Inst.		
	$\mathcal{A}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{A}$	$\mathcal{F}$	$\mathcal{P}$
LLaVA-NeXT-7B	58.5	63.6	26.7	58.9	61.8	4.1
Llama-3.2-11B-Vision	54.7	61.8	22.1	54.1	47.4	6.3
LLaVA-OneVision-7B	68.1	68.6	42.3	40.4	40.0	20.5
Gemma-3-12B	66.9	72.7	47.1	64.6	39.9	20.4
Qwen2.5-VL-7B	69.5	82.8	43.2	58.7	63.9	14.2
InternVL3-8B	71.9	80.4	47.3	67.7	75.6	29.8
Gemini2.0-Flash	**84.5**	**86.7**	**75.3**	**76.0**	**84.7**	**67.3**

Table 4: The average performance of MLLMs.

Following. Furthermore, all MLLMs demonstrate suboptimal performance in Preference, which reflects existing models' shortcomings in preference alignment.

RQ2 What conclusions can be drawn from the analysis of the stratified difficulty? Certain models exhibit deficiencies in handling specific difficulty levels of tasks and constraints, as shown in Table 3. For instance, Gemini underperforms when following C3 constraints, while Qwen shows notable limitations in solving T2 tasks. We conducted a further analysis of the failure cases corresponding to performance degradation, as illustrated in Figure 4. Our findings reveal that Qwen exhibits hallucination in recognizing objects' quantities, colors, and spatial relationships. Meanwhile, Gemini demonstrates difficulties in adhering to word-count-related instructions, frequently generating excessive content beyond the specified requirements. These findings underscore the necessity of graded classification for evaluating multimodal instruction-following capabilities.

RQ3 Are the models' abilities to "solve tasks" and to "follow constraints" equivalent to each other? We analyze the correlation between Accuracy and Following of MLLMs at different task levels as constraint difficulty increases, as illustrated in Figure 5. For instance, Gemini shows a correlation coefficient of 0.41 between $\mathcal{A}$ and $\mathcal{F}$ in Level 1 tasks. While its accuracy remains stable with increasing constraint difficulty, abnormal patterns emerge in constraint compliance. Notably, LlaVA-NeXT exhibits a negative correlation between $\mathcal{A}$ and $\mathcal{F}$ in Level 2 tasks. These findings further validate the necessity of our separate evaluation framework for Accuracy and Following.

RQ4 How do models perform when handling multi-modal instructions involving multiple tasks and constraints? We investigate the performance of the model in processing instructions with multiple tasks and constraints, as demonstrated in Table 4. As the number of tasks and constraints increases, the model shows a performance degradation in all three evaluation metrics, with $\mathcal{P}$ showing the most pronounced decline. In contrast, $\mathcal{F}$ remains relatively stable,

Model	MMIFEvol Single-Task $\mathcal{A}$	MMIFEvol Single-Task $\mathcal{F}$	MMIFEvol Single-Task $\mathcal{P}$	MMIFEvol Multi-Tasks $\mathcal{A}$	MMIFEvol Multi-Tasks $\mathcal{F}$	MMIFEvol Multi-Tasks $\mathcal{P}$	MIA-Bench Total	MM-IFEval Accuracy
LLaVA-NeXT-7B	58.5	63.6	26.7	58.9	61.8	4.1	69.4	36.4
+ Single-Inst.	65.4 (+6.9)	81.9 (+18.3)	29.9 (+3.2)	64.6 (+5.7)	65.2 (+3.4)	8.3 (+4.2)	73.5 (+4.1)	36.9 (+0.5)
+ Multi-Inst.	76.2 (+17.7)	88.9 (+25.3)	44.0 (+17.3)	64.6 (+5.7)	71.6 (+9.8)	28.8 (+24.7)	76.2 (+6.8)	41.5 (+5.1)
Qwen2.5-VL-7B	67.2	67.1	33.2	71.4	72.4	12.2	83.8	42.5
+ Single-Inst.	78.6 (+11.4)	93.1 (+26.0)	50.5 (+17.3)	74.9 (+3.5)	73.8 (+1.4)	16.9 (+4.7)	88.3 (+4.5)	43.2 (+0.7)
+ Multi-Inst.	79.7 (+12.5)	95.8 (+28.7)	54.8 (+21.6)	81.5(+10.1)	92.8 (+20.4)	28.1 (+15.9)	90.7 (+6.9)	45.6 (+3.1)
InternVL3-8B	71.9	80.4	47.3	78.6	85.5	23.1	84.5	44.1
+ Single-Inst.	77.8 (+5.9)	94.9 (+14.5)	52.8 (+5.5)	80.5 (+1.9)	85.5 (+0.0)	24.8 (+1.7)	87.1 (+2.6)	45.5 (+1.4)
+ Multi-Inst.	80.7 (+8.8)	95.8 (+15.4)	57.3 (+10.0)	81.5(+2.9)	94.2 (+8.7)	30.4 (+7.3)	93.6 (+9.1)	47.8 (+3.7)

Table 5: The performance of the model after SFT under the MMIFEvol synthetic instructions. Additionally, the models demonstrated improved performance on other multimodal instruction-following benchmarks.

Qwen2.5-VL-7B

Image	Instruction	GT	Response	A	F	P
	How many giraffe?	7	3	0	100	0
	Answer in JSON format: Where's the truck?	{ "direction": front, }	{ "bbox_2d":[140, 123, ...], }	0	100	0
	Describe as fairy tale: What color is the bus?	This bus loves to wear yellow robe ...	Buses wear blue clothes ...	0	25	0

Gemini2.0-Flash

Image	Instruction	GT	Response	A	F	P
	Color, 3 words.	Bright, warm, yellow	orange vivid full	100	100	0
	Color, 5 words.	Red, yellow, green, blue, white	Green orange yellow red	75	25	0
	Text, 1 word.	Fisketorvet	Shopping Center	0	0	0

Figure 4: Bad cases that occurred during the evaluation of Qwen2.5-VL-7B and Gemini2.0-Flash. Qwen demonstrated hallucination in T2 tasks (physical information recognition), while Gemini2.0 Flash tended to generate incorrect or redundant content when adhering to C3-type constraints (output length restrictions).

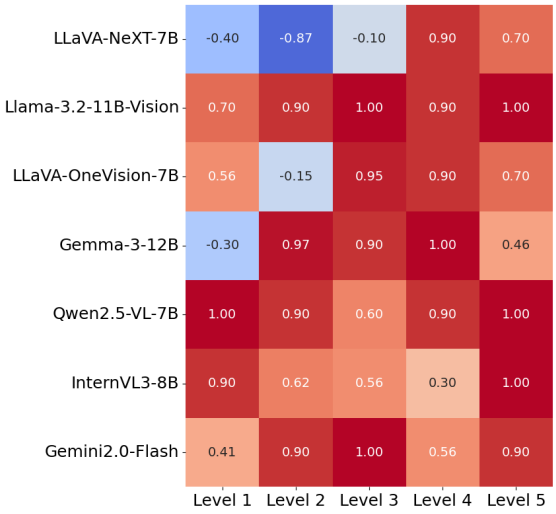

Figure 5: The Spearman rank correlation coefficient between $\mathcal{A}$ and $\mathcal{F}$. The model's ability to "solve tasks" and "follow constraints" is related, but not exactly consistent.

suggesting that the model retains its ability to sequentially process tasks and adhere to user requirements even under complex instruction scenarios.

RQ5 Can MMIFEvol improves the instruction-following ability of MLLM? We distilled 52k instruction responses from Qwen2.5-VL-72B and performed supervised fine-tuning (SFT), as shown in Table 5. Models after SFT demonstrate significantly enhanced instruction-following capabilities, while also showing improved performance on other benchmark tests. We compare two distillation approaches: single-task instruction and multi-task instruction distillation. The results indicate that multi-task instruction data contributes more substantially to model capability enhancement, underscoring the continued importance of instruction fine-tuning based on realistic, complex instructions.

Conclusion

This paper contributes to multimodal instruction following and benchmarking. We propose MMIFEvol, a framework that synthesizes and evolves instructions from a single image based on two dimensions: tasks and constraints. In addition, we design three comprehensive and fine-grained metrics to evaluate the capability of mainstream models. Experimental results prove that existing MLLMs still face challenges when executing complex instructions. We hope that this work will inspire future research to decouple task solving from constraint following and to develop more systematic method for multimodal instructions.

Acknowledgements

Thanks for the kind suggestions and support from Huawei Large Model Data Technology Lab.

References

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

Bai, T.; Liang, H.; Wan, B.; Xu, Y.; Li, X.; Li, S.; Yang, L.; Li, B.; Wang, Y.; Cui, B.; et al. 2024. A survey of multimodal large language model from a data-centric perspective. *arXiv preprint arXiv:2405.16640*.

Bitton, Y.; Bansal, H.; Hessel, J.; Shao, R.; Zhu, W.; Awadalla, A.; Gardner, J.; Taori, R.; and Schmidt, L. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv preprint arXiv:2308.06595*.

Cao, M.; Lam, A.; Duan, H.; Liu, H.; Zhang, S.; and Chen, K. 2024. Compassjudger-1: All-in-one judge model helps model evaluation and evolution. *arXiv preprint arXiv:2410.16256*.

Chen, G. H.; Chen, S.; Zhang, R.; Chen, J.; Wu, X.; Zhang, Z.; Chen, Z.; Li, J.; Wan, X.; and Wang, B. 2024. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*.

Chen, S.; Guo, X.; Li, Y.; Zhang, T.; Lin, M.; Kuang, D.; Zhang, Y.; Ming, L.; Zhang, F.; Wang, Y.; et al. 2025. Ocean-ocr: Towards general ocr application via a vision-language model. *arXiv preprint arXiv:2501.15558*.

Ding, S.; Wu, S.; Zhao, X.; Zang, Y.; Duan, H.; Dong, X.; Zhang, P.; Cao, Y.; Lin, D.; and Wang, J. 2025. Mmifengine: Towards multimodal instruction following. *arXiv preprint arXiv:2504.07957*.

Dong, R.; Han, C.; Peng, Y.; Qi, Z.; Ge, Z.; Yang, J.; Zhao, L.; Sun, J.; Zhou, H.; Wei, H.; et al. 2023. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*.

Feng, T.; Zhai, Y.; Yang, J.; Liang, J.; Fan, D.-P.; Zhang, J.; Shao, L.; and Tao, D. 2022. Ic9600: A benchmark dataset for automatic image complexity assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 8577–8593.

Forehand, M. 2010. Bloom's taxonomy. *Emerging perspectives on learning, teaching, and technology*, 41(4): 47–56.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Grill-Spector, K.; and Malach, R. 2004. The human visual cortex. *Annu. Rev. Neurosci.*, 27(1): 649–677.

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

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jiang, Y.; Wang, Y.; Zeng, X.; Zhong, W.; Li, L.; Mi, F.; Shang, L.; Jiang, X.; Liu, Q.; and Wang, W. 2023. Followbench: A multi-level fine-grained constraints following benchmark for large language models. *arXiv preprint arXiv:2310.20410*.

Lei, X.; Gomez, L.; Bai, H. Y.; and Bashivan, P. 2024. iwisdm: Assessing instruction following in multimodal models at scale. *arXiv preprint arXiv:2406.14343*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Li, M.; Zhang, Y.; Li, Z.; Chen, J.; Chen, L.; Cheng, N.; Wang, J.; Zhou, T.; and Xiao, J. 2023. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 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*, 740–755. Springer.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, J.; Huang, X.; Zheng, J.; Liu, B.; Wang, J.; Yoshie, O.; Liu, Y.; and Li, H. 2024. MM-Instruct: Generated Visual Instructions for Large Multimodal Model Alignment. *arXiv preprint arXiv:2406.19736*.

Luo, R.; Zhang, H.; Chen, L.; Lin, T.-E.; Liu, X.; Wu, Y.; Yang, M.; Wang, M.; Zeng, P.; Gao, L.; et al. 2024. Mmevol: Empowering multimodal large language models with evol-instruct. *arXiv preprint arXiv:2409.05840*.

Qian, Y.; Ye, H.; Fauconnier, J.-P.; Grasch, P.; Yang, Y.; and Gan, Z. 2024. Mia-bench: Towards better instruction following evaluation of multimodal llms. *arXiv preprint arXiv:2407.01509*.

Raichle, M. E.; and Mintun, M. A. 2006. Brain work and brain imaging. *Annu. Rev. Neurosci.*, 29(1): 449–476.

Ren, Q.; Zeng, J.; He, Q.; Liang, J.; Xiao, Y.; Zhou, W.; Sun, Z.; and Yu, F. 2025. Step-by-Step Mastery: Enhancing Soft Constraint Following Ability of Large Language Models. *arXiv preprint arXiv:2501.04945*.

Su, Y.; Lan, T.; Li, H.; Xu, J.; Wang, Y.; and Cai, D. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Tenenbaum, J. M. 1971. *Accommodation in computer vision*. Stanford University.

Wu, J.; Gan, W.; Chen, Z.; Wan, S.; and Yu, P. S. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, 2247–2256. IEEE.

Xu, G.; Jin, P.; Hao, L.; Song, Y.; Sun, L.; and Yuan, L. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Yang, X.; Wu, W.; Feng, S.; Wang, M.; Wang, D.; Li, Y.; Sun, Q.; Zhang, Y.; Fu, X.; and Poria, S. 2025. MM-InstructEval: Zero-shot evaluation of (Multimodal) Large Language Models on multimodal reasoning tasks. *Information Fusion*, 103204.

Zhang, L.; Cui, Q.; Zhao, B.; and Yang, C. 2025. Oasis: One image is all you need for multimodal instruction data synthesis. *arXiv preprint arXiv:2503.08741*.

Zhao, Y.; Huang, J.; Hu, J.; Wang, X.; Mao, Y.; Zhang, D.; Jiang, Z.; Wu, Z.; Ai, B.; Wang, A.; et al. 2025. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 29733–29735.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.