MM-IFEngine: Towards Multimodal Instruction Following

Shengyuan Ding[1,2*], **Shenxi Wu**[1,2*], **Xiangyu Zhao**[2,3], **Yuhang Zang**[2✉],
Haodong Duan[2], **Xiaoyi Dong**[2], **Pan Zhang**[2], **Yuhang Cao**[2], **Dahua Lin**[2,4,5], **Jiaqi Wang**[2✉]
[1]Fudan University [2]Shanghai AI Laboratory
[3]Shanghai Jiaotong University [4]The Chinese University of Hong Kong [5]CPII under InnoHK

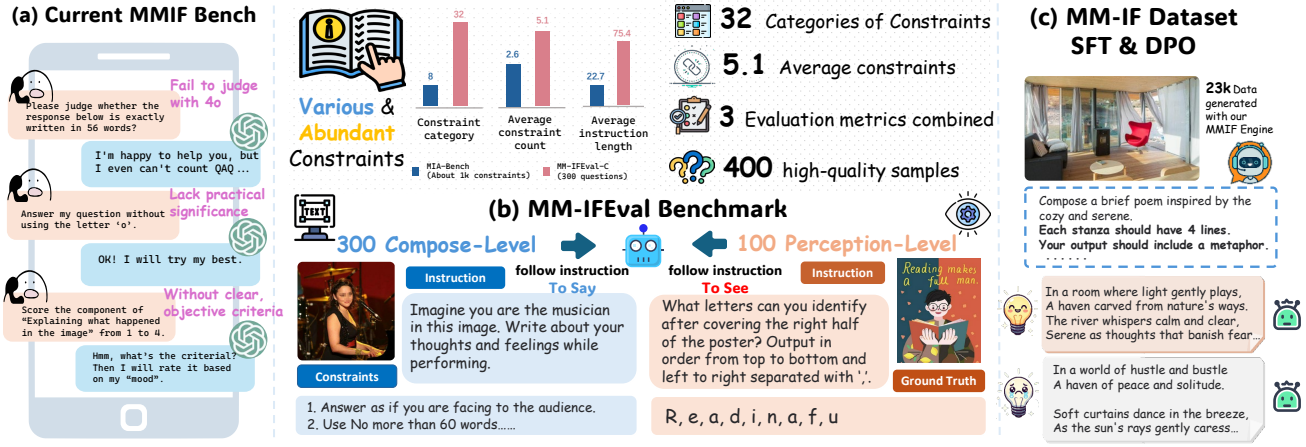

Figure 1. **(a)** Limitations of existing Multimodal Instruction Following (IF) benchmarks. **(b)** Overview of the MM-IFEval benchmark, which significantly surpasses existing benchmarks in terms of constraint diversity, quantity, and instruction complexity. Our benchmark consists of Compose-Level (C-Level) problems that impose constraints on model outputs (e.g., format requirements, keyword limits) and Perception-Level (P-Level) problems that require reasoning about specific visual elements in images. **(c)** Our MM-IFEngine generates a large-scale, diverse training dataset suitable for both Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO).

Abstract

The Instruction Following (IF) ability measures how well Multi-modal Large Language Models (MLLMs) understand exactly what users are telling them and whether they are doing it right. Existing multimodal instruction following training data is scarce, the benchmarks are simple with atomic instructions, and the evaluation strategies are imprecise for tasks demanding exact output constraints. To address this, we present MM-IFEngine, an effective pipeline to generate high-quality image-instruction pairs. Our MM-IFEngine pipeline yields large-scale, diverse, and high-quality training data MM-IFInstruct-23k, which is suitable for Supervised Fine-Tuning (SFT) and extended as MM-IFDPO-23k for Direct Preference Optimization (DPO). We further introduce MM-IFEval, a challenging and diverse multi-modal instruction-following benchmark that includes (1) both compose-level constraints for output responses and perception-level constraints tied to the input images, and (2) a comprehensive evaluation pipeline incorporating both rule-based assessment and judge model. We conduct SFT and DPO experiments and demonstrate that fine-tuning MLLMs on MM-IFInstruct-23k and MM-IFDPO-23k achieves notable gains on various IF benchmarks, such as MM-IFEval (+10.2%), MIA (+7.6%), and IFEval (+12.3%). We have fully open-sourced the datasets (both SFT and DPO), evaluation code and training scripts at https://github.com/SYuan03/MM-IFEngine.

1. Introduction

Instruction Following (IF) is a fundamental ability in Large Language Models (LLMs) [14, 27, 35, 53, 57] and Multimodal Large Language Models (MLLMs) [2, 34], which involves accurately interpreting and executing user-provided instructions. This ability is crucial for deploying models in real-world applications where users expect precise and context-aware responses, such as code generation [44], visual question answering [17], robots [38],

* Equal contribution. ✉ Corresponding authors.

and creative content creation [58]. For instance, in a VQA scenario, when a user asks an MLLM `what is the object and how do I use it, return the object name and the usage instructions in a JSON format`, accurate IF ensures the model provides a response like {`object`: '`hammer`', '`usage`': '`use it to drive nails`'} instead of the plain text.

Achieving precise IF in multimodal, diverse, and open-ended environments presents significant challenges for both *model training* and *benchmark evaluation*. One significant limitation is the scarcity of high-quality IF training data to train open-source MLLMs. In addition, current multimodal IF benchmarks [2, 34] merely have simple, atomic instructions, and the constraints are weakly correlated with visual content (see Fig. 1 **(a)**). Consequently, existing benchmarks lack the diversity required for real-world applications, leading to saturated results where nearly all models achieve over 80%. Furthermore, the evaluation method in existing benchmarks often relies on LLM-as-a-judge [56], which is imprecise for instructions demanding exact output constraints, such as word counts. Therefore, the combination of *limited training data*, *simple benchmarks*, and *imprecise evaluation strategy* strongly restricts the progress of current MLLMs in IF.

To address the lack of high-quality IF training data and challenging benchmarks, we propose **MM-IFEngine**, an effective pipeline for generating high-quality image-instruction pairs. MM-IFEngine collects diverse image sources, including natural scenes, UI interfaces, diagrams, charts, and mathematical problems. We then employ a structured approach using a predefined set of 16 task descriptions and 32 constraints to guide the LLM in crafting tailored instructions for each image. Using MM-IFEngine, we generated a comprehensive dataset of image-instruction pairs, collected responses from open-source MLLMs, and applied rigorous post-processing to retain only high-quality instruction-answer pairs, thus constructing **MM-IFInstruct-23k** for Supervised Fine-Tuning (SFT). We also generate negative responses by selectively removing constraints from the original data, constructing the preference dataset **MM-IFDPO-23k** for preference optimization algorithms such as Direct Preference Optimization (DPO) [36].

To facilitate the evaluation of multimodal IF, we present **MM-IFEval**, a benchmark comprising 400 challenging problems with diverse compose-level and perception-level instructions. MM-IFEval is derived from the images and instructions generated by MM-IFEngine with human-labeled annotations. As presented in Fig. 1 **(b)**, our MM-IFEval has the following three distinctive features: (1) **Diverse Instruction Types**: MM-IFEval has 32 distinct constraints, ensuring a wide range of instruction complexities and surpassing the scope of prior benchmarks. (2) **Hybrid Evaluation**: we use a hybrid strategy including both rule-based verification and judge model. For subjective instructions (e.g., mimicking tone), we design a *comparative* judgment for precise evaluation. Specifically, a control output is generated without the constraint, and the LLM judge compares both outputs for precise evaluation. (3) **Challenging**: the leading proprietary model (GPT-4o at 64.6%) and open-source model (Qwen2-VL-72B at 50.8%) demonstrating substantial room for improvement on our benchmark, highlights a significant opportunity for improvement in multimodal instruction following.

We further demonstrate that fine-tuning MLLMs on either MM-IFInstruct-23k or MM-IFDPO-23k consistently boosts the performance of MLLMs on instruction following benchmarks, without compromising their original capabilities on other Visual Question Answering (VQA) benchmarks. Specifically, fine-tuning Qwen2-VL-7B on MM-IFDPO-23k with the DPO results in performance gains of 10.2%, 7.6%, and 12.3% on MM-IFInstruct-23k, MIA-Bench [34], and IFEval [57], respectively.

Our contributions include: **(1)** a MM-IFEngine pipeline for generating multimodal constraint-rich image-instruction pairs; **(2)** a large-scale training dataset MM-IFInstruct-23k and preference optimization dataset MM-IFDPO-23k derived from MM-IFEngine; **(3)** a challenging multimodal instruction following benchmark MM-IFEval with diverse constraints and comprehensive evaluation approaches; and **(4)** empirical evidence showing significant performance gains on both our MM-IFEval and existing benchmarks when training MLLMs on MM-IFInstruct-23k via SFT and MM-IFDPO-23k via DPO.

2. Related Work

Instruction Following in LLMs. Various benchmarks and training approaches have been proposed to make Large Language Models (LLMs) better align with human instructions. While existing Instruction Following (IF) benchmarks like [14, 35, 53, 57] all aim to evaluate instruction following, they differ significantly in their *dataset construction pipelines*, driven by their unique constraint taxonomies. CFBench [53], for instance, constructs its dataset using a combination of taxonomic and statistical methodologies to establish comprehensive constraints. This divergence extends to their *evaluation strategies*. For example, InFoBench [35] adopts a strategy of decomposing complex instructions into simpler assessment standards. Beyond benchmarks, various training approaches aim to enhance LLMs' instruction-following capabilities [29, 44], including in-context learning [58] and preference optimization [54]. However, he aforementioned research is limited to the text modality, whereas our work focuses on multi-modal instruction following with vision inputs.

Instruction Following Benchmarks in MLLMs. Numerous benchmarks [18] have been proposed to evaluate di-

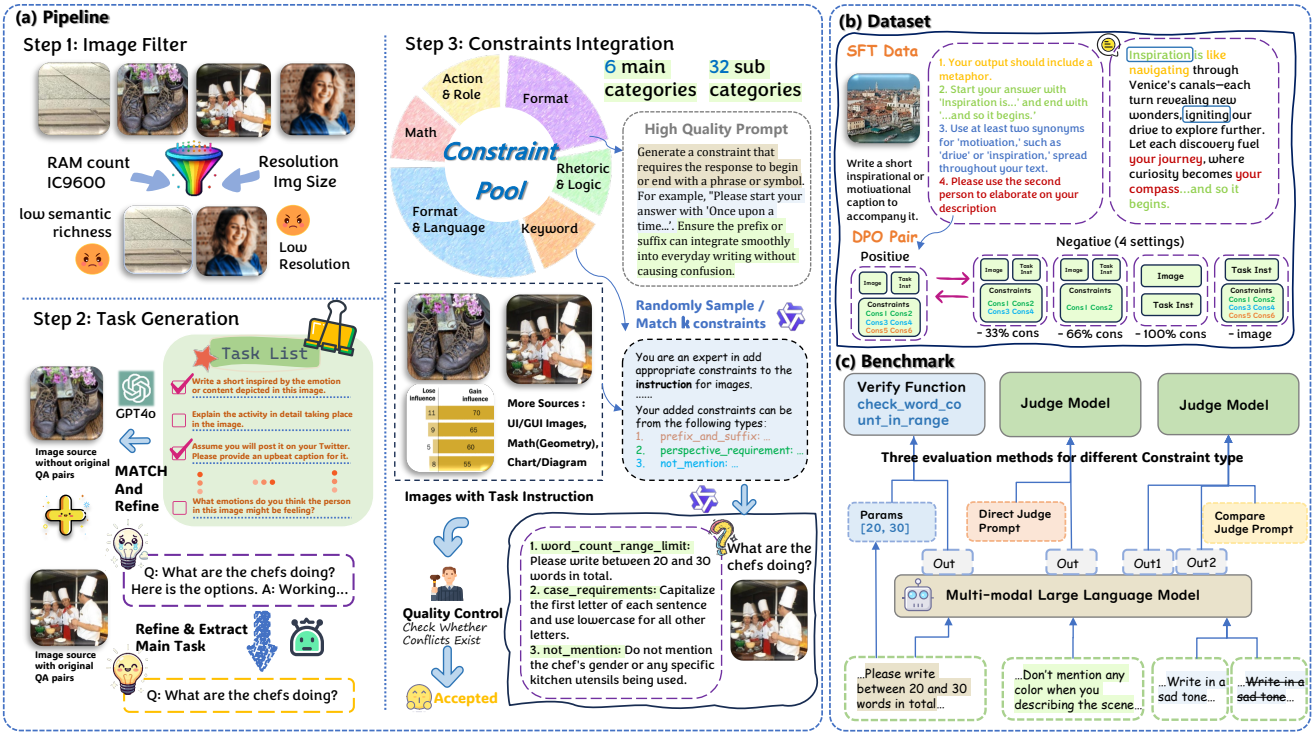

Figure 2. **Overall pipeline of MM-IFEngine**. Part **(a)** demonstrates the three-stage workflow of our engine: (1) Image filter; (2) Task generation using GPT-4o for images without QA pairs and instruct refinement for existing annotations; and (3) Constraints integration incorporating 6 main categories and 32 subcategories, ensuring compatibility between constraints and tasks. MM-IFEngine is employed to generate SFT and DPO training datasets and MM-IFEval benchmark, as shown in part **(b)** and **(c)**. MM-IFEval implements three evaluation metrics combining rule-based verification functions and a judge model to ensure accurate assessment.

verse capabilities of Multi-modal Large Language Models (MLLMs), including general knowledge [5, 24, 48, 50], document understanding [15, 25, 30], perception [43, 52], multi-image comprehension [26, 39, 40], and instruction following (IF) [2, 34]. MIA-Bench [34] and VisIT-Bench [2] are representative IF benchmarks that employ GPT-4 [32] for question generation and evaluation. In contrast to existing IF benchmarks, our MM-IFEval introduces significant improvements in diversity (32 constraint categories covering compositional and perceptual aspects), difficulty (averaging 5.1 constraints per question), and evaluation precision (using both judge models and rule-based verification).

Instruction Tuning Data for MLLMs. Recent advancements in multi-modal instruction tuning data aim to improve cross-modal alignment and increase the variety of tasks handled by MLLMs [4, 8, 20, 26, 45, 46, 51]. For example, some previous works [3, 4, 23] build synthetic instruction tuning data generated using GPT-4V [33], enabling open-source MLLMs to achieve performance comparable to proprietary models across multiple benchmarks. However, existing instruction tuning data are mainly designed for general knowledge or visual perception, and data for improving the IF abilities is scarce. The scarcity of training data for enhancing IF abilities motivated the development of our MM-IFEngine pipeline.

3. MM-IFEngine

We employ the MM-IFEngine pipeline to generate image-instruction pairs, which are the foundation for creating instruction tuning data and our benchmark. As shown in Fig. 2 **(a)**, the pipeline is composed of three main stages: (1) image filtering, where we systematically select a diverse set of images from multiple sources to ensure broad coverage of visual content; (2) task generation, in which we either synthesize novel tasks tailored to the selected images or refine existing instruction templates to better align with the image content; and (3) constraint integration, where high-quality, constraint-aware instructions are generated for images that initially lack associated annotated guidance, thereby enhancing the richness and precision of the dataset.

3.1. Image Filter

Our image filtering strategy selects only high-quality images by removing those with low resolution or limited semantic richness. For unannotated pure image datasets (e.g., CC3M

[37]), we prioritize natural scene images. Rich semantic content in these images enables the creation of more comprehensive and insightful QA pairs, which is crucial for designing diverse and complex instruction following tasks. We use the IC9600 and RAM metric proposed in the previous method [55] to select the images that have rich semantic content.

Furthermore, we analyze existing annotated datasets, such as ALLaVA [3]. Our analysis reveals that some images suffer from low resolution, making them inadequate for the instruction-following task. Given our intention to design more intricate and varied instruction following tasks based on this data, we filter out data items containing low-quality images.

3.2. Task Generation

Image Source without Original QA Pairs. For image datasets lacking original annotated task instructions (e.g., CC3M [37]), we first design appropriate task instructions for the data items. We first develop a series of task instructions tailored to the data items. These instructions are crafted to elicit long-form responses that can be subsequently modified or refined using various constraints, for instance, *Provide a detailed analysis of the image, including the setting, characters, and notable objects.* The final task pool $\mathcal{P}_T$ comprises a total of 16 distinct tasks, with further details available in Appendix A.1.2.

Given the task pool $\mathcal{P}_T$, we randomly select k tasks as examples of task types for each image I. We then prompt a powerful language model $\mathcal{M}$ (e.g., GPT-4o) to generate an appropriate task list T_l that aligns with the image content. The process is formulated as:

$$\{T_l^*\} = \mathcal{M}(I, T_e) \tag{1}$$

where $T_e = \{T_1, T_2, \ldots, T_k\}$ and each $T_i \in \mathcal{P}_T$. The model $\mathcal{M}$ is tasked with either choosing relevant tasks from T_e or supplementing reasonable tasks to construct the appropriate task list T_l^*, ensuring that all tasks in T_l^* are in line with the image content. After generating the T_l^*, a sampling step is incorporated to guarantee task diversity. For each image, tasks are sampled. This sampling process is crucial as it enriches the variety of tasks associated with each image.

Image Source with QA Pairs. In the case of image datasets that have QA pairs (e.g., ALLaVA [3]), we adopt certain strategies for processing the original question annotations. We choose ALLaVA as the primary dataset for this type of image source due to its rich and diverse image content, which is accompanied by a variety of task types. First, we conduct an analysis of the original question annotations. We find that some of the questions are accompanied by some few-shot examples. Additionally, some questions in ALLaVA have options in their original annotations, which are not suitable for our instruction-following task. Since we need to incorporate certain constraints into the original instructions in the subsequent steps, we use regular expressions and length limits to filter the questions in ALLaVA. Specifically, we select those questions that do not have few-shot examples associated with them. Mathematically, if we let Q be the set of all questions in ALLaVA, Q_{fs} be the subset of questions with few-shot examples, and Q_{op} be the subset of questions with options. We aim to find the subset Q_s of questions that satisfy the conditions:

$$Q_s = \{q \in Q | q \notin Q_{fs} \wedge q \notin Q_{op}\} \tag{2}$$

where the filtering based on the absence of few-shot examples and options is achieved using regular expressions and length limits. Then, we get the expected T^* in our filter Q_s set for the images.

3.3. Constraints Integration

Constraints Pool ($\mathcal{P}_C$) We use *instruction* to refer to the entire textual input, which in our paper can generally be viewed as a composition of *a task instruction* and *multiple constraints instruction*. Tasks and constraints are rich and diverse, with a certain complexity in our work. All the constraints in our work can be further classified into six major categories, each with its own unique characteristics and applications: Text Length Requirements, Mathematical Requirements, Language & Formatting Requirements, Rhetoric & Logic Requirements, Action Requirements, and Keyword Requirements. Please refer to the Appendix Fig. 5 for more details of all the constraints.

Given the constraints pool $\mathcal{P}_C$ and task instructions, a straightforward approach for composing full instruction is to first set several constraints for each constraint type and then randomly select one constraint from some of the types to compose the constraint list, and finally concatenate the constraint list with the task instruction to form the full instruction. But this direct method has two problems: (1) The constraints are not diverse enough, which may not be able to fully evaluate the ability of the model. (2) The contradiction between the constraints and also between the constraints and the task instruction may exist. For the first problem, an LLM is employed to generate concrete content of constraint instruction for the specific constraint type in our method. In order to avoid the generated content being too divergent or hard to control its difficulty, we carefully design some cases or requirements of details that needed to be paid attention to when generating the content for each constraint type (Appendix A.1.1). For the second problem, we also use a powerful LLM to help keep the correlation of constraints with its instruction and filter out those that cause total contradiction. Finally, we prompt an LLM to check whether the constraints and the task instruction are compatible and filter out those failing to pass the check. Our method not only ensures the compatibility of constraints and instructions but

also enriches the diversity of constraints.

In our actual practice process, we find that although we prompt the LLM to select appropriate constraints that should be compatible with the task instruction and other constraints, the generated constraints still have some contradiction with the task instruction, especially on those existing datasets with various kinds of annotations. The reason is that these datasets are designed for overall question-answering tasks, and the question(or named task instruction) tends to be contradictory with the constraints, which are mostly compatible with those tasks of creating or answering in non-short form. So, we decouple the selection and generation steps for this type of data source. Specifically, we first select the constraints from the constraints pool $\mathcal{P}_C$ and then provide the selected mostly compatible constraints to the LLM to select secondly and generate final constraints. But for image datasets without original QA pairs, in other words, for which we generate task instructions for them using $\mathcal{P}_T$, we directly sample k constraint types for the LLM to generate concrete content because they are mostly compatible with the pre-designed task instruction. The uniform process is formulated as:

$$C_l^* = \mathcal{L}(C_s, T^*), C_f^* = \mathcal{V}(C_l^*, T^*) \tag{3}$$

where $\mathcal{T}^*$ is the task applicable to the image. The model $\mathcal{L}$ is tasked with both choosing appropriate constraint types from C_s again and generating concrete constraints for some of them, whose output is a list of concrete constraint descriptions. To ensure that the generated constraints remain compatible with the given task instruction T^*, we employ a final validation step using another LLM process, denoted as $\mathcal{V}$. This validation function checks whether each constraint in C_l^* aligns with T^* and filters out those that contradict or do not fit the task instruction. The resulting set of fully verified and compatible constraints is represented as C_f^*.

MM-IFInstruct-23k Construction. By applying the MM-IFEngine pipeline, we construct the MM-IFInstruct-23k dataset, which contains 23k high-quality multi-modal instruction-following training data. We first take an analysis of the performance of the current open-source MLLMs and proprietary MLLMs on several benchmarks [25, 34], and find that for instruction-following capability, the most powerful open-source MLLM like InternVL2.5-78B-MPO [42] is nearly equivalent to GPT-4o, and the performance on general VQA benchmarks are even higher then GPT-4o. Thus, we use InternVL2.5-78B-MPO to generate responses for our MM-IFInstruct-23k dataset. Despite its capabilities, the InternVL2.5-78B-MPO model encounters difficulties in ensuring 100% compliance with our constraints, a challenge attributed to the complexity, number, and comprehensiveness. Consequently, we implement a post-processing stage to filter out responses that do not meet the specified criteria. Acknowledging that achieving perfect constraint adherence might be challenging even for human annotators on this task, we set a practical accuracy threshold of 80%. Finally, our MM-IFInstruct-23k comprises 23k data items, with 16k constructed from the training set of CC3M, 6k from ALLaVA, and 4k from the training set of MultiUI, Geo170k[12] and ChartQA[31]. We show the distribution of constraints number of MM-IFInstruct-23k in Fig. 3.

MM-IFDPO-23k Construction. To comprehensively explore and make full use of our high-quality data, we also utilize MM-IFEngine to construct MM-IFDPO-23k, a preference dataset comprising chosen and rejected samples suitable for Direct Preference Optimization (DPO) [36]. Our high-quality data can be directly employed as the chosen samples. Regarding rejected samples, we opt to utilize Qwen2-VL-7B-Instruct to answer the variant of the question for generating rejected pairs. Specifically, we have four distinct settings for generating negative pairs, which mainly differ in the input to Qwen2-VL-7B-Instruct. These settings include (1) With image, but randomly remove one-third of the number of constraints in the prompt; (2) With image, but randomly remove two-thirds of the number of constraints in the prompt; (3) With image, but randomly remove all the constraints in the prompt; and (4) Full prompt, but without the image; We use these four types of input to feed into Qwen2-VL-7B-Instruct model, and collect the rejected responses to construct the MM-IFDPO-23k.

4. MM-IFEval

Existing benchmarks for multi-modal instruction following are scarce. The majority focus on simple and atomic instructions, resulting in performance saturation across models. To address this limitation, we introduce **MM-IFEval**, a human-annotated, comprehensive, and challenging benchmark designed for evaluating multi-modal instruction following.

4.1. MM-IFEval Construction

To construct the MM-IFEval, we first use our MM-IFEngine to generate the question-answer (QA) pairs for images. The generated instructions may inherently contain potential conflicts. Consequently, human annotation remains critical for constructing this benchmark, as human annotators possess the cognitive capacity for comprehensive assessment of these complex situations. After the human annotation, we further use an extra post-processing step that prompts the LLMs to double-check and mitigate the occurrence of constraint conflicts as much as possible. Finally, we construct the MM-IFEval bench of 400 questions, 300 of which are *compose-level* open-ended questions and 100 *perception-level* questions with ground truth.

Diverse Constraints. With 32 distinct constraint categories and an average of 5.1 constraints per question, MM-IFEval presents a more challenging evaluation task compared to earlier benchmarks (e.g., [34], which has 8 categories and 2.6 average constraints per question). Furthermore, our bench-

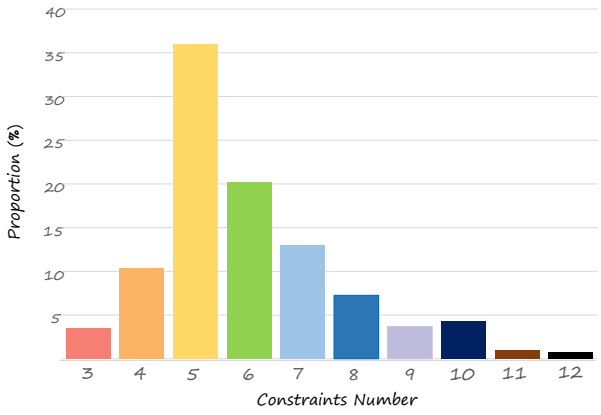

Figure 3. **Constraint Quantity Distribution in MM-IFInstruct-23k**. Our MM-IFInstruct-23k exhibits systematic variation in constraint complexity, with each sample containing 3-12 constraints per instruction.

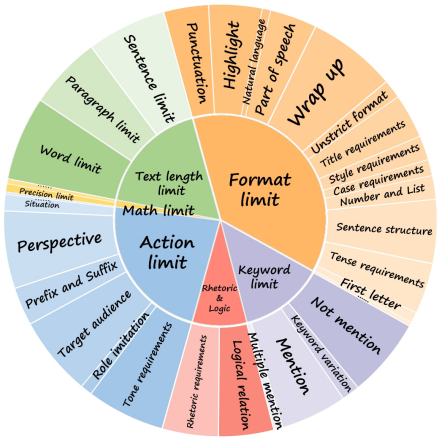

Figure 4. **Constraint Category Distribution in Compose-Level Problems of MM-IFEval**. This part comprises six primary constraint categories with 32 subcategories, forming a multi-level taxonomy for instruction-following evaluation.

mark incorporates essential constraints such as "Output in JSON format", which is prevalent and practical in real-world scenarios, a feature not found in previous multi-modal instruction following benchmarks.

Compose-level and Perception-level Questions. *Compose-level* questions involve textual constraints, while *perception-level* questions require greater visual perception ability to solve. The perception-level questions incorporate a variety of image sources, such as natural scenes, user interfaces, diagrams, table charts, and mathematical expressions, which we believe are representative of real-world applications. Please refer to the Appendix for examples of compose-level and perception-level questions.

4.2. Hybrid Evaluation

Current multi-modal instruction following benchmarks often rely solely on GPT-4o for evaluation. However, accurately assessing certain constraints, such as numerical conditions (e.g., '`output in 200 words`', '`Answer in 5 paragraphs`', '`Use the word 'cat' in the answer twice`'), remains challenging even for GPT-4o. In contrast, verifiable functions like string matching offer greater precision than judge models for such constraints. To address this, we propose a hybrid evaluation strategy (see Fig. 2(c)) that employs three methods, including both rule-based Verification and judge models for more robust and precise evaluation. We also conducted a human–LLM agreement study on 30 sampled items (162 constraints), achieving a consistency score of 90.74%, further validating the benchmark's reliability.

(1) **Rule-based Verification.** For constraints that adhere to a fixed format and involve specific content that can be objectively verified—yet remain challenging for an LLM to assess accurately—we employ a rule-based approach. Specifically, we design a set of predefined functions for different constraint types. The LLM is first prompted to extract the relevant parameters, denoted as *Params*, from the constraint description. When evaluating a constraint that falls within the scope of our rule-based framework, we use *Params* and the model's output as inputs to the predefined function to determine compliance.

(2) **LLM-based Direct Judgment.** This method is primarily used for evaluating constraints that can be easily and unambiguously verified based on the model's output. It is applicable to constraints where correctness is straightforward to determine, such as those requiring the inclusion of specific words or phrases. For instance, a constraint like "Use the word 'inspiration' or its synonyms at least twice in the response" does not follow a strict format and cannot be assessed using a rule-based approach. Instead, we directly leverage an LLM to determine whether the constraint is satisfied.

(3) **LLM-based Comparative Judgment.** Some constraints, particularly those related to tone, style, or role-playing, are difficult to evaluate directly. To improve judgment accuracy, we adopt a comparative approach. Specifically, we generate a second model output using a nearly identical prompt but without the constraint under evaluation. The LLM-based evaluator is then provided with both outputs and asked to compare them, determining whether the model's response with the constraint in the prompt adheres more closely to the expected requirement.

5. Experiments

Benchmarks. We select the following benchmarks to demonstrate that models fine-tuned on MM-IFInstruct-23k

Table 1. **Main results on Instruction Following benchmarks**, including our proposed MM-IFEval, MIA-Bench [34], and IFEval [57]. The symbol ^{M} refers to multimodal benchmarks, and ^{T} denotes text-only benchmarks. We report both compose-level ("C") and perception-level ("P") for MM-IFEval, prompt-level accuracy ("Prompt.") and Inst-level accuracy ("Inst.") for IFEval, and the averaged results across all three benchmarks in the rightmost column.

Model	Parameter	MM-IFEval^{M} (ours)			MIA^{M}	IFEval^{T}			Avg.
		C	P	Avg.		Prompt.	Inst.	Avg.	
LLaVA-NeXT-7B [21]	7B	36.8	16.0	31.6	73.2	32.0	43.3	37.7	47.5
LLaVA-OneVision-Qwen2-7B-OV [16]	8B	37.4	24.0	34.0	84.5	43.3	54.8	49.0	55.8
InternVL2-8B [7]	8B	45.2	32.0	41.9	86.2	44.6	57.0	50.8	59.6
InternVL2.5-8B [6]	8B	49.6	36.0	46.2	88.5	52.2	62.4	57.3	64.0
LLaVA-NeXT-Llama3-8B [21]	8B	45.9	21.0	39.7	83.3	45.0	56.4	50.7	57.9
w. MM-IFInstruct-23k	-	59.3	19.0	49.2 +9.5	86.5 +3.2	50.8	61.8	56.3 +5.6	64.0 +6.1
w. MM-IFDPO-23k	-	58.7	21.0	49.3 +9.6	90.0 +6.7	64.5	73.7	69.1 +18.4	**69.5** +11.6
Qwen2-VL-7B-Instruct [41]	8B	42.7	40.0	42.0	80.5	42.4	52.5	47.4	56.6
w. MM-IFInstruct-23k	-	57.0	38.0	52.3 +10.3	87.7 +7.2	46.8	58.4	52.6 +5.2	64.2 +7.6
w. MM-IFDPO-23k	-	55.2	43.0	52.2 +10.2	88.1 +7.6	55.2	64.3	59.7 +12.3	**66.7** +10.1

Table 2. **Main results on VQA benchmarks**, including general knowledge (MMMU [50], MMBench [24], MMStar [5], MMT-Bench [48]), document understanding (AI2D [15], OCRBench [25]), Chat (MMVet [49]) and Hallusion (POPE [19]). Fine-tuning models on MM-IFDPO-23k achieve comparable performance across these benchmarks.

Model	General				Document		Chat	Hallusion	Avg.
	MMMU_{val}	$\text{MMBench}_{\text{dev}}$	MMStar	$\text{MMT-Bench}_{\text{val}}$	AI2D	OCRBench	MMVet	POPE	
LLaVA-NeXT-Llama3-8B [21]	43.7	72.5	43.6	53.1	73.1	55.0	43.3	87.2	58.9
w. MM-IFInstruct-23k	45.8	69.3	44.2	53.3	71.2	55.3	46.3	88.8	59.3
w. MM-IFDPO-23k	44.1	72.1	43.7	53.1	72.3	56.7	43.9	86.8	59.1
Qwen2-VL-7B-Instruct [41]	53.9	81.0	60.8	63.2	82.9	86.7	63.3	86.3	72.3
w. MM-IFInstruct-23k	54.0	79.3	57.1	61.0	81.6	81.8	61.6	89.2	70.7
w. MM-IFDPO-23k	54.0	81.3	58.5	63.7	83.3	86.8	66.1	85.7	72.4

and MM-IFDPO-23k enhance instruction following without compromising performance on other VQA tasks: (1) **Instruction Following benchmarks**, including MIA-Bench [34], IFEval [57], and our proposed MM-IFEval. To be noted, IFEval is a language-only benchmark while others are both multi-modal benchmarks. (2) **VQA Benchmarks**, including MMMU [50], MMBench [24], MMStar [5], AI2D [15], OCRBench [25], MMVet [49], POPE [19] and MMT-Bench [48].

Implementation Details. We fine-tuned Qwen2-VL-7B-Instruct [41] and LLaVA-Next-Llama3-8B [21] using MM-IFInstruct-23k (SFT) and MM-IFDPO-23k (DPO). For the SFT phase, we used a batch size of 128 and a learning rate of 1e-5. For the DPO phase, we used a learning rate of 5e-7 with the batch size of 16. We implemented our training pipeline with the help of LLaMA-Factory and evaluation pipeline under VLMEvalkit [10].

5.1. Results about MM-IFInstruct-23k and MM-IFDPO-23k

Consistently Improvements on Instruction Following Benchmarks. As shown in Tab. 1, both MM-IFInstruct-23k and MM-IFDPO-23k significantly enhance the model's performance in instruction following benchmarks. Fine-tuning LLaVA-Next and Qwen2-VL on MM-IFInstruct-23k yielded significant averaging performance gains of 6.1% and 7.6% points, respectively. Furthermore, applying DPO with MM-IFDPO-23k also led to notable improvements for LLaVA-Next and Qwen2-VL, with average gains of 11.6% and 10.1% points. Such improvements demonstrate the effectiveness of MM-IFEngine in constructing high-quality training data.

Comparable Results on VQA Benchmarks. To show that fine-tuning on MM-IFInstruct-23k and MM-IFDPO-23k improves instruction following without degrading performance on other VQA tasks, we analyzed model performance on other widely used benchmarks, as detailed in Tab. 2. Results indicate that models fine-tuning with MM-IFInstruct-23k and MM-IFDPO-23k demonstrate comparable performance across these benchmarks.

SFT vs DPO. As evidenced by Tab. 1 and Tab. 2, DPO using MM-IFDPO-23k significantly surpasses SFT on MM-IFInstruct-23k. This is likely due to negative samples of DPO, which are essential for training models to respect constraints, particularly in our data with multiple and diverse constraints. Additionally, the Kullback–Leibler (KL) divergence in DPO preserves the model's generalization, as demonstrated in Tab. 2.

Table 3. **Evaluation of various MLLMs on MM-IFEval**. We report the accuracy of easy and difficult problems and the average accuracy across all problems. The C-Level and P-Level refer to the compose-level and perception-level problems, respectively. The **best** performance in each section is highlighted in **bold**.

Model	Param	C-Level	P-Level	Avg.
Proprietary MLLMs				
Claude-3.5V-Sonnet [1]	-	67.5	**44.0**	61.7
GPT-4o-mini [13]	-	70.4	40.0	62.8
GPT-4o (20240806) [13]	-	**71.5**	**44.0**	**64.6**
Open-Source MLLMs				
LLaVA-NeXT-7B [21]	7B	36.8	16.0	31.6
LLaVA-OneVision-Qwen2-7b-OV [16]	8B	37.4	24.0	34.0
MiniCPM-V-2.6 [47]	8B	39.2	32.0	37.4
InternVL2-8B [7]	8B	45.2	32.0	41.9
InternVL2-40B [7]	40B	48.0	36.0	45.0
InternVL2.5-8B [6]	8B	49.6	36.0	46.2
InternVL2.5-26B [6]	8B	53.5	32.0	48.1
Qwen2-VL-72B-Instruct [41]	72B	53.4	**43.0**	50.8
LLaVA-NeXT-Llama3-8B [21]	8B	45.9	21.0	39.7
+ MM-IFDPO-23k	-	**58.7**	21.0	49.3
Qwen2-VL-7B-Instruct [41]	8B	42.7	40.0	42.0
+ MM-IFDPO-23k	-	55.2	**43.0**	**52.2**

5.2. Leaderboard of MM-IFEval

We present the performance comparison results of various MLLMs on our MM-IFEval in Tab. 3, including both proprietary MLLMs such as GPT-4o [13] and Claude-3.5 [1] and open-source MLLMs such as LLaVA-Next [21], LLaVA-OneVision [16], InternVL [6, 7], and Qwen2-VL [41].

MM-IFEval is Challenging. Results on Tab. 3 demonstrate that multimodal instruction following is still a challenging and unsolved task for current MLLMs, specifically for the perception-level problems. The propriety models GPT-4o and Claude-3.5V-Sonnet establish top-tier average performance with scores of 64.6 and 61.7, respectively. The leading open-source MLLM, Qwen2-VL-72B merely achieves an overall accuracy of 50.8. We attribute the performance gap between proprietary and open-source models to the scarcity of high-quality open-source training data for instruction following. As a result of our MM-IFDPO-23k, Qwen2-VL-7B fine-tuned via our optimized DPO approach achieves a score of 52.2, demonstrating a 24.3% relative improvement over its baseline (42.0), and even surpasses the larger Qwen2VL-72B model. We hope our MM-IFEval benchmark motivates further exploration into improving MLLM instruction-following. Additional benchmark examples from MM-IFEval, including images and constraint-based instructions for both compose-level and perception-level tasks, are provided in the Appendix.

5.3. Ablation Studies

Ablation Studies on Different DPO Settings. In Tab. 4, we present an ablation study on various strategies for constructing pairwise preference data for Direct Preference Optimization (DPO). These strategies primarily include: (1) generating rejected responses by randomly removing constraints from the instruction (second to fourth rows), and (2) prompting MLLMs without providing image inputs to generate rejected responses (bottom row).

Table 4. **Ablation studies across different DPO settings**, including randomly deleting constraints (second row to fourth row) or prompting MLLMs without images (bottom row) to generate negative responses. Avg. refers to the average score of three IF benchmarks.

Model	MM-IFEval	MIA	IFEval	Avg.
Qwen2-VL-7B-Instruct	42.0	80.5	47.4	56.6
+ DPO (-33% cons)	51.5	**88.2**	57.9	65.8
+ DPO (-66% cons)	51.2	88.0	58.4	65.9
+ DPO (-100% cons)	**52.2**	88.1	**59.7**	**66.7**
+ DPO (w/o img)	48.4	86.9	54.7	63.4
LLaVA-NeXT-Llama3-8B	39.7	83.3	50.7	57.9
+ DPO (-33% cons)	**50.4**	87.2	64.3	67.3
+ DPO (-66% cons)	48.7	86.8	**69.7**	68.4
+ DPO (-100% cons)	49.3	**90.0**	69.1	**69.5**
+ DPO (w/o img)	44.7	85.9	64.8	65.2

We conduct experiments on both the Qwen2-VL-7B-Instruct and LLaVA-NeXT-Llama3-8B models. As shown in Tab. 4, all DPO variants exhibit strong robustness, consistently outperforming the baseline. Among the four evaluated strategies, removing 100% of the constraints to generate rejected responses achieves the best performance, whereas omitting image inputs yields the weakest performance. Furthermore, we observe a consistent trend: as the proportion of removed constraints increases from 33% to 100%, the performance of the resulting DPO models improves accordingly. This suggests that removing more constraints amplifies the semantic gap between preferred and rejected responses, thereby enhancing the effectiveness of contrastive learning during DPO training. Based on these findings, we adopt the 100%-constraint removal strategy as the default approach for constructing the DPO data in MM-IFDPO-23k.

6. Conclusion

This paper contributes to the field of multimodal instruction-following by exploring pipelines for training data collection and proposing a challenging benchmark. We present MM-IFEngine, a pipeline designed to generate image-instruction pairs, subsequently used to construct MM-IFInstruct-23k for SFT and MM-IFDPO-23k for DPO. We also analyze the limitations of existing multimodal instruction following benchmarks and propose MM-IFEval, a benchmark featuring diverse instruction types and a hybrid evaluation strategy that combines rule-based methods with an LLM-based judge. We hope this work inspires further research into improving the instruction-following ability of MLLM, a critical step towards unlocking their potential in real-world applications.

Acknowledgement

This work was supported by National Key R&D Program of China 2022ZD0161600, Shanghai Artificial Intelligence Laboratory, the Shanghai Postdoctoral Excellence Program (No.2023023), China Postdoctoral Science Fund (No.2024M751559, No.2025T180412), Hong Kong RGC TRS T41-603/20-R, the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK. Dahua Lin is a PI of CPII under the InnoHK.

References

[1] Anthropic. Claude 3.5 sonnet. 2024. 8

[2] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. VisIT-Bench: A benchmark for vision-language instruction following inspired by real-world use. In *NeurIPS, Datasets and Benchmarks*, 2023. 1, 2, 3

[3] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024. 3, 4, 2

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

[5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024. 3, 7

[6] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 7, 8

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 7, 8

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 3

[9] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*, pages 845–854, 2017. 2

[10] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024. 7

[11] Xinyu Fang, Zhijian Chen, Kai Lan, Shengyuan Ding, Yingji Liang, Xiangyu Zhao, Farong Wen, Zicheng Zhang, Guofeng Zhang, Haodong Duan, et al. Creation-mmbench: Assessing context-aware creative intelligence in mllm. *arXiv preprint arXiv:2503.14478*, 2025. 3

[12] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023. 5, 2

[13] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 8

[14] Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. Followbench: A multi-level fine-grained constraints following benchmark for large language models. In *ACL*, 2024. 1, 2

[15] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 3, 7

[16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 7, 8

[17] Huayang Li, Siheng Li, Deng Cai, Longyue Wang, Lemao Liu, Taro Watanabe, Yujiu Yang, and Shuming Shi. TextBind: Multi-turn interleaved multimodal instruction-following in the wild. In *ACL Findings*, 2024. 1

[18] Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024. 2

[19] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023. 7

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 3

[21] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 7, 8

[22] Junpeng Liu, Tianyue Ou, Yifan Song, Yuxiao Qu, Wai Lam, Chenyan Xiong, Wenhu Chen, Graham Neubig, and Xiang Yue. Harnessing webpage uis for text-rich visual understanding, 2024. 2

[23] Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, Yu Qiao, and Jifeng Dai. Mminstruct: a high-quality multi-modal instruction tuning dataset with extensive diversity. *Science China Information Sciences*, 67(12), 2024. 3

[24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He,

Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In *ECCV*, 2024. 3, 7

[25] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 2024. 3, 5, 7

[26] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. MMDU: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. In *NeurIPS Datasets and Benchmarks Track*, 2024. 3

[27] Renze Lou, Kai Zhang, and Wenpeng Yin. A comprehensive survey on instruction following. *arXiv preprint arXiv:2303.10475*, 2023. 1

[28] Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*, 2024. 3

[29] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*, 2023. 2

[30] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. MMLongBench-Doc: Benchmarking long-context document understanding with visualizations. In *NeurIPS Datasets and Benchmarks Track*, 2024. 3

[31] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 5, 2

[32] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. Accessed: 2025-02-23. 3

[33] OpenAI. GPT-4V(ision) System Card. 2023. Accessed: 2025-02-23. 3

[34] Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. MIA-Bench: Towards better instruction following evaluation of multimodal llms. In *ICLR*, 2025. 1, 2, 3, 5, 7

[35] Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. InFoBench: Evaluating instruction following ability in large language models. *arXiv preprint arXiv:2401.03601*, 2024. 1, 2

[36] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 2, 5

[37] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 4, 2

[38] Lucy Xiaoyang Shi, Brian Ichter, Michael Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, et al. Hi Robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025. 1

[39] Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context, 2024. 3

[40] Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding, 2024. 3

[41] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 7, 8

[42] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024. 5

[43] Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, et al. Videorope: What makes for good video rotary position embedding? *arXiv preprint arXiv:2502.05173*, 2025. 3

[44] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023. 1, 2

[45] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning, 2023. 3

[46] Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning, 2024. 3

[47] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 8

[48] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi, 2024. 3, 7

[49] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 7

[50] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren,

Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 3, 7

[51] Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 2025. 3

[52] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *IJCV*, 2025. 3

[53] Tao Zhang, Yanjun Shen, Wenjing Luo, Yan Zhang, Hao Liang, Fan Yang, Mingan Lin, Yujing Qiao, Weipeng Chen, Bin Cui, et al. CFBench: A comprehensive constraints-following benchmark for llms. *arXiv preprint arXiv:2408.01122*, 2024. 1, 2

[54] Xinghua Zhang, Haiyang Yu, Cheng Fu, Fei Huang, and Yongbin Li. Iopo: Empowering llms with complex instruction following via input-output preference optimization, 2024. 2

[55] Xiangyu Zhao, Shengyuan Ding, Zicheng Zhang, Haian Huang, Maosong Cao, Weiyun Wang, Jiaqi Wang, Xinyu Fang, Wenhai Wang, Guangtao Zhai, et al. Omnialign-v: Towards enhanced alignment of mllms with human preference. *arXiv preprint arXiv:2502.18411*, 2025. 4

[56] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS Datasets and Benchmarks Track*, 2023. 2

[57] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. 1, 2, 7

[58] Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. Controlled text generation with natural language instructions. In *ICML*, 2023. 2