Towards Realistic Conversational Multimodal Instruction Following

Kaiwen Tuo[1,2*] **Congcong Wang**[2*] **Shuai Dong**[3*] **Hongrui Wu**[4*]
Siyuan Wang[5] **Xuefeng Yin**[6] **Yuhang Cao**[2] **Nan Duan**[2] **Jiaqi Wang**[2,†]
[1]The Hong Kong University of Science and Technology [2]Joy Future Academy [3]Fudan University
[4]Stanford University [5]The Chinese University of Hong Kong [6]Tongji University
cfintuo@gmail.com, wjqdev@gmail.com

Abstract

Instruction following is a core capability of modern foundation models, underpinning assistant-style interaction, agentic workflows, visual understanding, and multimodal reasoning. Existing instruction-following data and methods, however, remain substantially misaligned with realistic user behavior: users often converse with models over multiple turns, provide interleaved images and videos, switch modalities, or implicitly expect earlier constraints to persist. To bridge this gap, we first identify and formalize **six core challenges** for realistic conversational multimodal instruction following: Constraint-Guided Visual Reasoning, Cross-Modal Context Transfer, Persistent Instruction Tracking, User-Aware Context Modeling, Iterative Grounded Response Refinement, and Visual Consistency Alignment. Driven by these challenges, we develop a multi-agent data construction pipeline that generates diverse, constraint-rich, image- and video-grounded conversations, followed by human verification. We further improve model behavior through a **rubric-guided multi-stage training pipeline** that decomposes complex instructions into atomic supervision signals and trains models from basic visual constraint following to advanced multi-turn conversational constraint satisfaction. Finally, we establish Multimodal MultiChallenge (MMMC) as a systematic benchmark for realistic conversational multimodal instruction following. The results show that realistic conversational multimodal instruction following remains challenging for frontier models, while targeted rubric-guided post-training improves constraint satisfaction and broadly preserves general multimodal capability.

1 Introduction

Instruction following (IF) [44, 64, 65] has become one of the defining interfaces between humans and frontier large language models [43, 24, 4]. A model that cannot reliably respect user constraints, for example, what evidence to use, what not to mention, how to format an answer [77, 29, 48, 28, 15], or which part of an image or video to inspect [16, 51, 27], is difficult to deploy in assistant and agent workflows [61]. While training paradigms like instruction tuning and reinforcement learning from human feedback (RLHF) [12, 46] serve as the cornerstone for this capability, current research still underspecifies the conditions under which users actually interact with multimodal assistants. Real users rarely query single-turn isolated prompts [7, 14]. Instead, they engage in dynamic conversations involving modality switching and iterative refinement, a reality that existing training data and benchmarks often fail to capture.

Within these realistic, dynamic interactions, the nature of constraints fundamentally changes, requiring models to coordinate IF with visual grounding [30], OCR [38], and long-context memory [59]. For

*Equal contribution. † Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison of representative instruction-following benchmarks across features, including modality, interaction setting, training pipeline, and main constraint type.

Benchmark	Modality	Turns	Visual-Centric	Cross-modal	Training pipeline	Main Constraint Type
IFEval [77]		Single	✗	✗	✗	Verifiable format/style
IFBench [48]		Single/Multi	✗	✗	✓	Verifiable out-of-domain constraints
Multi-IF [28]		Multi	✗	✗	✗	Multi-turn multilingual constraints
MultiChallenge [15]		Multi	✗	✗	✗	Realistic conversational challenges
MM-IFEval [16]		Single	✓	✗	✓	Compose/perception constraints
MMMT-IF [20]		Multi	✓	✗	✗	Global format constraints
IF-VidCap [34]		Single	✓	✗	✓	Caption-oriented constraints
MCIF [45]		Single	✓	✓	✗	Scientific-talk tasks
Audio MultiChallenge [25]		Multi	Audio-native	✗	✗	Spoken multi-turn challenges
MMMC (ours)		Single/Multi	✓	✓	✓	Multimodal conversational constraints

instance, image inputs introduce spatial regions and attribute-filtered selection [70, 32], while video inputs necessitate managing temporal windows and event-based reasoning [31, 3]. Multi-turn dialogue further compounds these challenges, as models must remember global formats, transfer referents across modalities, and preserve answer stability during selective editing [35]. In such scenarios, instruction following is no longer merely about surface-form formatting. The model must jointly manage visual grounding, temporal localization, conversational memory, cross-modal transfer, selective editing, and self-consistency.

To bridge this gap and explicitly enhance model capabilities, we propose a systematic solution for training robust multimodal LLMs. We first identify and formalize six core challenge categories that capture the failure modes of current models: (i) **Constraint-Guided Visual Reasoning (CGVR)**, (ii) **Cross-Modal Context Transfer (CMCT)**, (iii) **Persistent Instruction Tracking (PIT)**, (iv) **User-Aware Context Modeling (UCM)**, (v) **Iterative Grounded Response Refinement (IGRR)**, and (vi) **Visual Consistency Alignment (VCA)**. Based on these categories, we develop a multi-agent data construction pipeline to generate diverse, constraint-rich, and human-verified conversational data. This pipeline provides the necessary data foundation to teach models how to navigate the intricate intersection of visual perception and conversational state tracking.

Equipped with this robust data foundation, we subsequently tackle model enhancement by introducing a rubric-guided multi-stage post-training framework. By decomposing complex instructions into atomic, human-verified rubrics, we provide fine-grained supervision signals that guide the model through a curriculum-style learning process, progressing from basic visual grounding to advanced conversational constraint satisfaction. This approach enables the model to master complex IF tasks while broadly preserving general multimodal capability across the evaluated benchmarks. Finally, we synthesize these efforts into the Multimodal MultiChallenge (MMMC) benchmark, see Table 1. Our experiments show that while realistic multimodal IF remains a significant hurdle for frontier models, our targeted training recipe substantially improves constraint satisfaction and model reliability.

Our contributions are as follows:

- **Training & Data Pipeline:** We design a multi-agent data synthesis pipeline and a rubric-guided training framework that explicitly targets six core multimodal conversational challenges, significantly improving the model's ability to follow complex, multi-turn instructions.
- **Evaluation Benchmark:** We introduce MMMC, a benchmark specifically designed for realistic conversational multimodal instruction following across images and videos, featuring a human-verified rubric-decomposed evaluation protocol.
- **Experimental Analysis:** We provide a comprehensive evaluation of frontier models, demonstrating that our proposed training stages effectively mitigate common failure modes such as constraint omission and conversational state drift.

2 Related Work

Instruction Following in Large Language Models. Instruction following has become a fundamental capability of large language models, shaping how models interpret user intent and satisfy explicit constraints. Early work mainly improved this ability through instruction tuning and align-

ment [65, 44, 64, 61, 76, 12]. Recent benchmarks formulate instruction following as fine-grained constraint satisfaction. IFEval [77] established a reproducible protocol based on programmatically checkable requirements, including length, keyword, and formatting constraints. Subsequent benchmarks broaden this paradigm with hierarchical difficulty, compositional requirements, and multilingual settings [29, 53, 66, 74, 36]. More recent efforts, including IFBench [48], EIFBench [79], and AgentIF [50], shift the focus toward unseen verifiable constraints, extremely complex workflows, and agentic system prompts.

Multimodal Instruction-Following Data and Evaluation. Multimodal large language models extend instruction following beyond text-only prompts to visual, audio, and video inputs. Visual instruction tuning, represented by LLaVA, shows that synthetic image-instruction data can equip language models with basic visual chat abilities [37]. Subsequent benchmarks like VisIT-Bench [6] and MM-Vet [71] evaluate MLLMs on real-world vision-language instructions and integrated capabilities. More recent work makes multimodal instruction the evaluation target. MIA-Bench [51], MM-IFEngine [16], and VC-IFEval [27] study whether MLLMs can satisfy layered image-grounded instructions, specified answer patterns, and visual-centric constraints. Beyond static images, video [34] and audio [45] introduce temporal evidence, dynamic events, speech content, and long-context multimodal dependencies.

Multi-Turn Dialogue Instruction-Following Data. Real user interactions are rarely single-turn [35, 7, 14], as users often refine requests, refer back to previous context, and expect earlier constraints to remain active. Therefore, recent benchmarks begin to examine instruction following in multi-turn conversations. MultiChallenge [15] identifies realistic conversational failure modes and evaluates them with instance-level rubrics. In multimodal settings, MMMT-IF [20] studies image-based multi-turn instruction following with global instructions distributed across the conversation. Nevertheless, existing work still lacks a unified benchmark for realistic conversations involving images, videos, modality switching, visual memory, temporal constraints, and evidence-grounded self-coherence.

Post-Training for Instruction Following. Instruction-following ability is commonly improved through supervised finetuning [37, 16], preference optimization, and reinforcement learning from feedback [17], where a key challenge lies in providing precise and informative feedback signals. InstructGPT demonstrated that supervised demonstrations, reward modeling, and PPO-based RLHF can better align models with user intent [44, 56]. For precisely verifiable constraints, RLVR has become appealing because reward functions can be computed automatically from format checking [48]. VerIF further investigates verification engineering for instruction-following RL by combining rule-based checks with LLM-based verification, and constructs VerInstruct with associated verification signals [46]. Our work follows this broader trend but focuses on multimodal and conversational constraints, leveraging fine-grained rubrics to improve instruction-following capabilities.

3 Conversational Multimodal Instruction Following

To improve multimodal instruction following, we formalize conversational constraint satisfaction and introduce a taxonomy of six core challenges, as illustrated in Figure 1.

3.1 Problem Formalization

Traditional instruction following (IF) focuses on single-turn prompts or image queries. We define a multi-turn dialogue trajectory as $\mathcal{T} = \{(q_1, \mathcal{M}_1, a_1), \ldots, (q_n, \mathcal{M}_n, a_n)\}$, comprising textual queries q_t, optional multimodal inputs $\mathcal{M}_t$, and responses a_t.

The constraints governing response a_n are rarely confined to query q_n. Instead, a complex set of active constraints C_{active} is derived from and distributed across the dialogue history $\mathcal{T}_{<n}$. To successfully follow instructions, the model must execute a joint function of visual perception, contextual retrieval, and constraint satisfaction. It must ground its answer in the specific visual evidence required by $\mathcal{M}_n$, while adhering to the cumulative logical, stylistic, and formatting rules established in $\mathcal{T}_{<n}$. The inability of current models to manage this intersection of visual reasoning and long-context state tracking motivates our taxonomy of challenges.

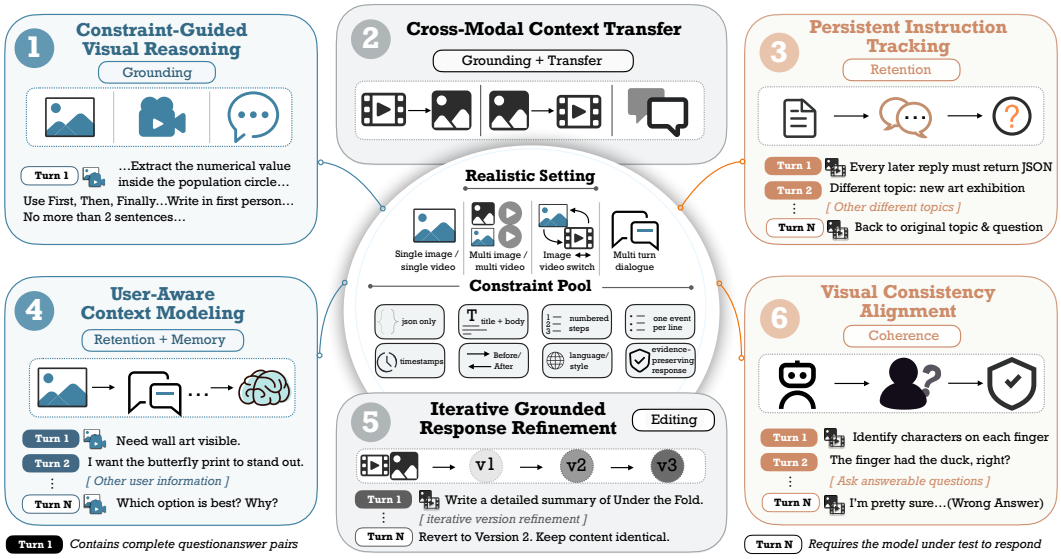


Figure 1: Illustration of the six core challenge families and the constraint pool in our conversational multimodal instruction following taxonomy, with realistic settings and constraint pools.

3.2 The Six Challenge Families

Based on the realistic settings depicted in Figure 1, we identify six categories of challenges that commonly emerge when users interact with multimodal LLMs. These categories are paired with a diverse constraint pool to rigorously test model capabilities.

Constraint-Guided Visual Reasoning. When users interact with multimodal assistants, they often do not ask the model to describe an entire image or video. Instead, they specify which part of the visual input should be used, which object should be followed, or which temporal segment should be considered. Constraint-Guided Visual Reasoning is designed to evaluate this basic yet crucial capability of following visually grounded restrictions. For a video, the user may ask the model to focus on a certain time window or track a specified object. For an image, the user may restrict the answer to a spatial region or objects with certain attributes. This category represents the atomic form of conversational multimodal instruction following.

Cross-Modal Context Transfer. In realistic conversations, users frequently switch between images and videos while expecting the assistant to maintain the same goal, referent, or response policy [45, 55]. Cross-Modal Context Transfer evaluates whether a model can preserve such information across modality changes. For instance, a user might establish a rule, or a user goal, and subsequently upload an image, and ask a follow-up question that implicitly depends on the earlier context.

Persistent Instruction Tracking. In practical interactions with LLMs, human users commonly specify a general instruction early in a conversation and expect the assistant to follow it throughout later turns. However, existing models often struggle to retain such instructions over extended dialogues, and this challenge becomes more pronounced when the conversation involves new visual inputs. Persistent Instruction Tracking evaluates whether multimodal models can maintain such global constraints after the dialogue has moved on to intermediate topics or different media. In such cases, users rarely restate the rule; therefore, the model must reliably retrieve the earlier instruction from the conversational history and apply it to the current visual task.

User-Aware Context Modeling. Beyond explicit instructions, users interacting with LLMs frequently reveal preferences, goals, roles, and task-specific needs throughout a conversation [67, 75]. User-Aware Context Modeling evaluates whether a model can recall and apply such information when it becomes implicitly relevant to a later multimodal request. For example, earlier turns may indicate that the user prefers high-level summaries, is preparing a tutorial, cares about safety-relevant observations, or wants to focus on a particular category of visual evidence. When a new image or video is introduced later, the model must apply this implicitly gathered profile without being explicitly reminded.

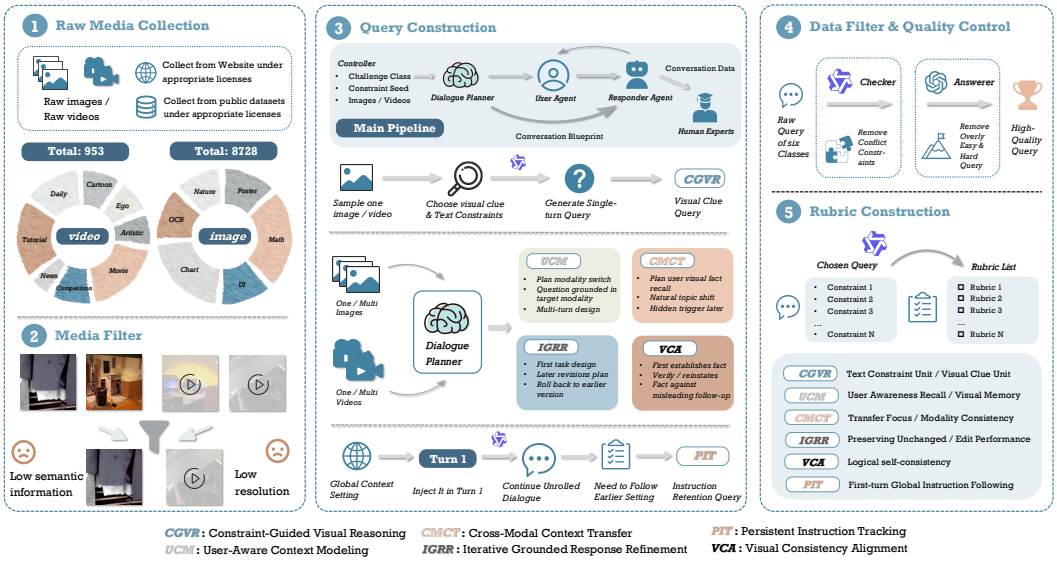


Figure 2: Overview of our multi-agent data synthesis pipeline, covering media collection as well as the detailed workflow for query and rubric annotation.

Iterative Grounded Response Refinement. Another common use case of multimodal LLMs is collaborative editing. Users may ask the model to write an answer based on an image or video, then revise only part of it, remove certain details, change the observation scope, or return to an earlier version. Iterative Grounded Response Refinement evaluates whether a model can maintain and modify visual answers across multiple turns without unnecessarily rewriting stable content. LLMs must resolve which version is being referred to and preserve both the visual facts and the constraints that should remain untouched. This category therefore tests instruction following, contextual reference resolution, and the visual faithfulness of LLMs simultaneously.

Visual Consistency Alignment. LLMs often exhibit sycophancy, tending to align their responses with user beliefs or preferences even when doing so conflicts with truthfulness or prior context [47, 58]. Visual Consistency Alignment evaluates whether a model can maintain reasonable consistency with its own visually grounded previous responses. The model is expected to respond coherently: it should not blindly agree with the user when a new claim conflicts with the visual evidence or with a previously verified answer.

4 Methodology

To explicitly equip models with the capabilities required to overcome the conversational and visual challenges defined in Section 3, we introduce an integrated methodology, a multi-agent data synthesis pipeline coupled with a rubric-guided post-training framework.

4.1 Multi-Agent Data Synthesis Pipeline

We construct MMMC through a multi-agent-based data construction framework, which is designed to produce realistic conversational multimodal instruction-following examples while ensuring that each instance is visually grounded and constraint-rich, see Figure 2. We first collect diverse images and videos from public or licensed sources, covering diverse domains including real-world scenarios [26, 60], synthetic media [54], GUIs [33, 69], and structured documents/charts [63, 40, 41, 18, 2, 39]. We use Qwen3.5-397B-A17B as the Media Filter to assess the semantic content and visual clarity of the collected media. We also construct a constraint pool containing a diverse set of visual-centric constraints and textual output constraints, detailed in Appendix B and Tables 7 and 8.

For each media group, we assign a challenge family to the planning agent (**Planner**), and ask it to design the full dialogue trajectory according to the given media. The visual constraints and textual output constraints to be incorporated into the dialogue are directly sampled from our pre-constructed constraint pool. The planner further checks whether the sampled constraints conflict with each

other and resamples them when necessary. Given the resulting plan, two dialogue generation agents (**Dialoguer**) comprise two roles with separate system prompts. They write natural user queries and generate simulated assistant responses from LLMs. The controller attaches each media file to its designated user turn. The assistant's response to the final user turn is intentionally withheld so that it can be produced by the model under evaluation. Finally, human annotators verify the generated examples. We inspect the media, dialogue, and reference answers to ensure that the requested task is answerable from the visual evidence. This human-in-the-loop verification step improves the reliability of the benchmark.

4.2 Rubric-Decomposed Representation

In such a complex setting, accurately evaluating whether a model response follows all instructions becomes challenging. Each instance may contain multiple constraints, and the judge model must often reason over a long multi-turn dialogue history. We therefore annotate each instance with detailed atomic rubrics, which provide explicit evaluation criteria for judging the correctness.

Formally, a MMMC instance is a multimodal conversational example: $x_i = (\mathcal{C}_i, q_i, \mathcal{M}_i, a_i^\star)$, where $\mathcal{C}_i$ denotes the dialogue context, q_i denotes the current user request, $\mathcal{M}_i$ denotes the optional multimodal input containing zero or more images and videos, and $a_i^\star$ denotes the reference assistant response. During evaluation, the model is given only the observable input $z_i = (\mathcal{C}_i, q_i, \mathcal{M}_i)$ and is required to generate a response $\hat{a}_i$. The reference response $a_i^\star$ is not provided to the model.

Each instance is associated with a set of atomic rubric items: $\mathcal{R}_i = \{r_{i,1}, r_{i,2}, \ldots, r_{i,K_i}\}$, where K_i is the number of rubric items for instance i. Each rubric item checks one specific requirement of the target response. Given the model response $\hat{a}_i$, each rubric item $r_{i,k}$ is evaluated by either a deterministic checker or a rubric-based judge model, producing a binary satisfaction score:

$$s_{i,k} = \begin{cases} 1, & \text{if } \hat{a}_i \text{ satisfies } r_{i,k} \text{ under } z_i, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

We report two evaluation metrics. The first is the **Rubric-level Score**, which measures the average fraction of satisfied atomic rubrics across all test instances:

$$S_{\text{rubric}} = \frac{1}{N} \sum_{i=1}^{N} S_i = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{K_i} \sum_{k=1}^{K_i} s_{i,k}, \tag{2}$$

where N denotes the number of test instances.

The second is the **Prompt-level Score**, which is a stricter metric that counts an instance as correct only if all rubric items are satisfied:

$$S_{\text{prompt}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\sum_{k=1}^{K_i} s_{i,k} = K_i \right]. \tag{3}$$

The **Prompt-level Score** reflects whether a model fully follows all constraints in an instance, while the **Rubric-level Score** provides a more fine-grained view of partial instruction satisfaction and enables detailed error analysis.

The prompt-level score above averages over all test instances and corresponds to the micro average. To give each challenge family equal weight despite differences in sample counts, we also report the macro Average of prompt-level scores across the six challenge families:

$$S_{\text{prompt}}^{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} \frac{1}{N_c} \sum_{i \in \mathcal{D}_c} \mathbb{I}\left[\sum_{k=1}^{K_i} s_{i,k} = K_i \right], \tag{4}$$

where $C = 6$, $\mathcal{D}_c$ is the set of test instances in challenge family c, and $N_c = |\mathcal{D}_c|$. Unless otherwise specified, the overall averages in our MMMC result tables refer to this macro-averaged prompt-level score. Scores are reported as percentages.

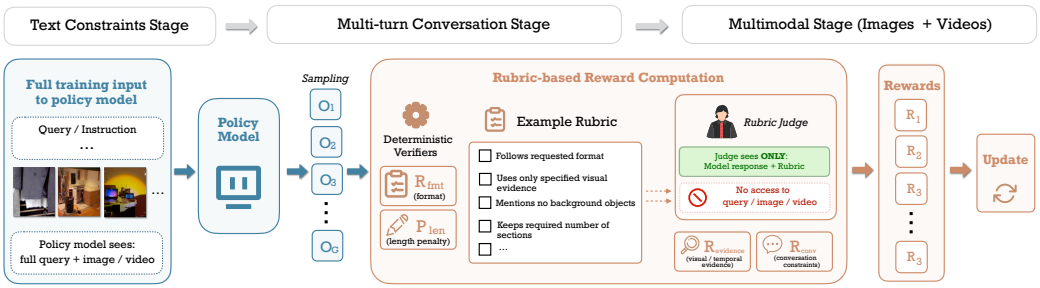


Figure 3: Illustration of MMMC Training Pipeline. The pipeline features a three-stage curriculum from text to full multimodal instructions. Rewards are computed by a Rubric Judge using atomic rubrics, strictly without accessing raw multimodal inputs.

4.3 Rubric-Based Reinforcement Learning

Baseline. We conduct baseline experiments with GRPO [57] on our single-turn CGVR data. We find that, as constraints become more complex, conventional reward-model judgments become increasingly unreliable. The inclusion of image and video inputs further raises the requirements for the reward model, since it must assess whether the response satisfies visually grounded constraints. Meanwhile, we observe that CGVR examples remain non-trivial for Qwen3-VL-8B, suggesting that even single-turn multimodal instruction following poses a meaningful challenge.

To address these issues, we **(i)** construct fine-grained rubrics for the training set, so that the reward model only needs to judge a model response against the corresponding rubric, without directly processing the original image or video. **(ii)** We adopt a multi-stage rubric-based RL training pipeline, which progressively activates the model's conversational multimodal instruction-following ability.

To progressively teach the model to follow multimodal instructions in a more robust manner, we design a multi-stage training paradigm tailored to instruction following, see Figure 3. Using our data construction framework, we first build text-only data centered on text constraints, which activates the model's basic ability to satisfy format, count, and other verifiable requirements. We then construct a set of text-only multi-turn dialogue data in forms analogous to PIT, UCM, IGRR, and VCA, enabling the model to adapt to context-dependent instruction following. Finally, we construct image and video data for the CGVR category, which bridges precise text-constraint satisfaction with visually grounded instruction following. Through training at each stage, the model steadily improves its ability to follow complex instructions.

For rubric-RL, we sample model responses and compute a reward from deterministic verifiers and rubric judges:

$$R(\hat{a}) = \lambda_{\text{fmt}} R_{\text{fmt}}(\hat{a}) + \lambda_{\text{rub}} \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}[J(\hat{a}, r_k) = \text{YES}], \tag{5}$$

where R_{fmt} denotes the rule-based format reward. Given a response $\hat{a}$ and a list of K atomic rubric items $\{r_k\}_{k=1}^{K}$, we use a judge model J to evaluate each item independently. The judge returns a binary verdict for each rubric item: the response receives 1 point if it satisfies r_k and 0 otherwise. The final rubric reward is computed as the average score over all rubric items. In our training setting, the judge model only observes the model response and the rubric item, without access to the original image, video, or full user query. This design decomposes complex multimodal instruction following into fine-grained verifiable checks and reduces the burden on the reward model to directly reason over raw multimodal inputs.

5 Experiments

Models and Baselines. For training, we instantiate our rubric-guided multi-stage post-training pipeline on the Qwen3-VL series [5], and report results on both Qwen3-VL-8B and Qwen3-VL-4B-Instruct to examine the effectiveness of our method across different model scales. For evaluation, we benchmark a broad set of representative multimodal large language models, including frontier proprietary models such as GPT-5.4 [43], Gemini 3.1 Pro [24], Doubao Seed 2.0 Pro [9], GPT-4o

[42], Gemini 3 Flash [23], and Doubao Seed 1.6 [8]. We further evaluate competitive open-source models, including Kimi-K2.5 [62], Qwen3-VL-32B, and Qwen3.5-9B [5]. This model suite allows us to assess both the absolute difficulty of realistic conversational multimodal instruction following and the effectiveness of our targeted training pipeline under controlled model-family comparisons. During training, we adopt GRPO [57] as our main training method, and further include the data and training recipe from VerIF [46] as a baseline for comparison.

Benchmarks. We evaluate on MMMC and public instruction-following benchmarks including IFEval [77], MM-IFEval [16], IFBench [48], MIA-Bench [51], and MultiChallenge [15]. To test whether instruction-following training harms general ability, we further evaluate our model on a suite of text-only and multimodal benchmarks, covering mathematical reasoning [1, 52], general knowledge [72, 73, 11], visual reasoning [10, 68], and video-domain understanding [22, 21, 78, 49, 13]. We conduct multimodal benchmark evaluations using our customized version of VLMEvalKit [19]. For rubric-based evaluation on MMMC, we use Qwen3.5-397B-A17B as the judge model.

Training Details. During training, we use gpt-oss-20B as the rubric judge for computational efficiency. Both judges evaluate each atomic rubric independently, receiving only the candidate response and one rubric item, without the original media, user query, or dialogue history. Rubric annotation uses the full visual input to encode the information needed for these checks. Deterministic rules verify mechanically checkable constraints, while the judge evaluates the remaining constraint types. The stage-specific learning rates, epochs, and GRPO sampling sizes are reported in Appendix C.

5.1 The MMMC Benchmark

To evaluate models' instruction-following ability in realistic multimodal scenarios, we use our data construction framework to build a comprehensive and challenging benchmark, MMMC, consisting of representative examples from the six challenge families introduced in this work. Before query generation, we split all media at the source level to ensure that no media used in the benchmark appear in the training set. We also ensure that the training set contains no examples from other instruction-following benchmarks.

Table 2 reports model performance across the six MMMC challenge families. Our training method and data bring clear improvements on the benchmark, especially in the User-Aware Context Modeling and Cross-Modal Context Transfer categories, and substantially outperform other baselines.

5.2 Results on Public Instruction-Following Benchmarks

Table 3 evaluates whether training on MMMC transfers to existing IF benchmarks. We evaluate our model on a representative set of challenging text-only, multimodal, and multi-turn instruction-following benchmarks. Under our training framework, the model exhibits clear improvements in multimodal instruction following, with a gain of 8.08 points on MM-IFEval. Meanwhile, the model also shows a stronger ability to handle long-context conversational constraints. These results indicate that our training pipeline not only improves performance on MMMC but also transfers effectively to public instruction-following evaluations, outperforming the baseline training methods.

5.3 General Multimodal Capability

We evaluate whether instruction-following training preserves general task-solving capability on the text-only and multimodal benchmarks in Table 4 and the video benchmarks in Table 5. Overall, our training broadly preserves general capability across mathematical reasoning, general knowledge, visual reasoning, and video understanding benchmarks, with improvements on some tasks and modest declines on others.

Our training improves performance on several benchmarks, including WeMath, MMMU, MMMU-Pro, and Video-Count-Eval. These gains are associated with improved task completion, answer consistency, and selective visual grounding. In particular, reduced repetitive generation helps models complete mathematical solutions, while better consistency between intermediate reasoning and final answers prevents some answer-selection errors. Training with visual constraints may also help models focus on relevant regions, objects, and events, benefiting tasks that require targeted visual recognition or counting.

Table 2: Main results on MMMC. Scores are prompt-level scores. Macro Avg. denotes the macro average across the six challenge families, and Micro Avg. denotes the average across all 551 examples. For the trained models, we report mean scores over five evaluation runs. CGVR: Constraint-Guided Visual Reasoning; CMCT: Cross-Modal Context Transfer; PIT: Persistent Instruction Tracking; UCM: User-Aware Context Modeling; IGRR: Iterative Grounded Response Refinement; VCA: Visual Consistency Alignment.

Model	CGVR	CMCT	PIT	UCM	IGRR	VCA	Macro Avg.	Micro Avg.
Proprietary models								
GPT-5.4	13.04	31.82	44.21	58.06	41.59	37.50	37.70	39.93
Gemini 3.1 Pro	17.39	25.00	48.42	64.52	40.19	52.78	41.38	44.10
Doubao Seed 2.0 Pro	8.70	31.82	48.42	67.74	47.66	40.28	40.77	44.10
GPT-4o	17.39	20.45	41.05	45.16	43.93	17.36	30.89	33.58
Gemini 3 Flash	13.04	27.27	47.37	48.39	28.50	50.00	35.76	37.75
Doubao Seed 1.6	8.70	27.27	42.11	51.61	41.59	36.81	34.68	38.48
Open-source models								
Kimi-K2.5	4.35	20.45	47.37	67.74	48.60	43.06	38.59	43.92
Qwen3-VL-32B-Instruct	13.04	18.18	23.16	61.29	44.86	25.69	31.04	33.58
Qwen3.5-9B	8.70	18.18	21.05	45.16	42.52	23.61	26.54	30.67
Qwen3-VL-4B-Instruct	13.04	11.36	18.95	29.03	38.79	23.61	22.46	27.59
Our trained models								
Qwen3-VL-8B	13.04	7.27	18.32	41.29	39.72	21.53	23.53	27.66
+ GRPO	11.30 (-1.74)	13.18 (+5.91)	22.95 (+4.63)	46.45 (+5.16)	41.68 (+1.96)	25.97 (+4.44)	26.92 (+3.39)	31.07 (+3.41)
+ VerIF	13.04 (+0.00)	11.82 (+4.55)	18.74 (+0.42)	38.71 (-2.58)	40.28 (+0.56)	19.72 (-1.81)	23.72 (+0.19)	27.70 (+0.04)
+ MMMC-Train (Ours)	15.65 (+2.61)	20.45 (+13.18)	21.68 (+3.36)	56.77 (+15.48)	45.23 (+5.51)	31.39 (+9.86)	31.86 (+8.33)	34.99 (+7.33)
Qwen3-VL-4B-Instruct	13.04	10.91	19.16	30.32	39.25	22.64	22.55	27.59
+ GRPO	13.04 (+0.00)	10.91 (+0.00)	20.42 (+1.26)	40.00 (+9.68)	41.68 (+2.43)	25.56 (+2.92)	25.27 (+2.72)	30.05 (+2.46)
+ VerIF	13.04 (+0.00)	12.27 (+1.36)	20.42 (+1.26)	35.48 (+5.16)	40.37 (+1.12)	23.06 (+0.42)	24.11 (+1.56)	28.75 (+1.16)
+ MMMC-Train (Ours)	13.91 (+0.87)	31.36 (+20.45)	21.89 (+2.73)	52.90 (+22.58)	42.15 (+2.90)	29.58 (+6.94)	31.97 (+9.42)	33.94 (+6.35)

Table 3: Results on public instruction-following benchmarks. Average is the unweighted mean of MM-IFEval, IFBench, MIA-Bench, and MC Average.

Model	MM-IFEval	IFBench	MIA-Bench	MultiChallenge					Average
				IM	SC	IR	RVE	MC Average	
Qwen3-VL-8B	61.12	31.29	84.61	43.36	24.00	46.38	24.39	34.53	52.89
+ GRPO	63.62 (+2.50)	35.03 (+3.74)	85.98 (+1.37)	40.71 (-2.65)	20.00 (-4.00)	46.38 (+0.00)	26.83 (+2.44)	33.48 (-1.05)	54.53 (+1.64)
+ VerIF	65.59 (+4.47)	32.31 (+1.02)	86.18 (+1.57)	41.59 (-1.77)	24.00 (+0.00)	52.17 (+5.79)	21.95 (-2.44)	34.93 (+0.40)	54.75 (+1.87)
+ MMMC-Train (Ours)	69.20 (+8.08)	45.57 (+14.28)	88.28 (+3.67)	73.45 (+30.09)	36.00 (+12.00)	66.67 (+20.29)	19.51 (-4.88)	48.91 (+14.38)	62.99 (+10.10)
Qwen3-VL-4B-Instruct	60.19	34.35	84.09	41.59	16.00	39.13	26.83	30.89	52.38
+ GRPO	60.62 (+0.43)	34.35 (+0.00)	86.92 (+2.83)	40.71 (-0.88)	22.00 (+6.00)	46.38 (+7.25)	29.27 (+2.44)	34.59 (+3.70)	54.12 (+1.74)
+ VerIF	63.29 (+3.10)	30.61 (-3.74)	87.85 (+3.76)	38.94 (-2.65)	18.00 (+2.00)	44.93 (+5.80)	29.27 (+2.44)	32.78 (+1.89)	53.63 (+1.25)
+ MMMC-Train (Ours)	69.38 (+9.19)	42.17 (+7.82)	88.51 (+4.42)	61.95 (+20.36)	24.00 (+8.00)	49.28 (+10.15)	24.39 (-2.44)	39.90 (+9.01)	59.99 (+7.61)

These improvements are accompanied by modest declines on some benchmarks. Both model scales show small declines on MLVU-Dev and VCRBench. Nevertheless, the average score across the mathematical, general-knowledge, and visual-reasoning benchmarks increases by 1.61 points for the 8B model and 1.25 points for the 4B model, while the video benchmark average changes by $+0.81$ and -0.20 points, respectively. These mixed results suggest that our training broadly preserves general capability, with task-specific gains and regressions rather than uniform improvements across all benchmarks.

5.4 Ablation Study

Table 6 ablates training stages. We separately train the model on the constructed text-only data, image data, and video data, and report the effect of each stage. The results show that each multimodal type of training data can independently bring consistent improvements to the model, demonstrating the effectiveness of our staged data construction and training design.

Table 4: General multimodal capability evaluation across mathematical reasoning, general knowledge, and visual reasoning benchmarks.

Model / Stage	Math		General Knowledge			Visual Reasoning		Average
	AIME25	WeMath	MMMU	MMMU-Pro	SimpleVQA	MMStar	RealWorldQA	
Qwen3-VL-8B-Instruct	39.90	52.19	66.11	51.79	53.09	70.67	70.98	57.82
+ GRPO	39.17 (-0.73)	52.10 (-0.09)	66.44 (+0.33)	52.14 (+0.35)	52.69 (-0.40)	71.00 (+0.33)	70.46 (-0.52)	57.71 (-0.11)
+ VerIF	40.83 (+0.93)	51.52 (-0.67)	67.00 (+0.89)	51.71 (-0.08)	52.54 (-0.55)	70.73 (+0.06)	70.98 (+0.00)	57.90 (+0.08)
+ MMMC-Train (Ours)	41.67 (+1.77)	53.81 (+1.62)	67.56 (+1.45)	53.18 (+1.39)	54.07 (+0.98)	72.13 (+1.46)	73.59 (+2.61)	59.43 (+1.61)
Qwen3-VL-4B-Instruct	41.77	47.33	62.11	47.92	50.07	69.27	71.37	55.69
+ GRPO	42.50 (+0.73)	47.81 (+0.48)	62.78 (+0.67)	47.89 (-0.03)	50.72 (+0.65)	70.60 (+1.33)	71.90 (+0.53)	56.31 (+0.62)
+ VerIF	40.10 (-1.67)	47.24 (-0.09)	63.22 (+1.11)	48.09 (+0.17)	51.70 (+1.63)	70.47 (+1.20)	72.29 (+0.92)	56.16 (+0.47)
+ MMMC-Train (Ours)	40.00 (-1.77)	51.62 (+4.29)	62.33 (+0.22)	48.82 (+0.90)	51.56 (+1.49)	70.40 (+1.13)	73.86 (+2.49)	56.94 (+1.25)

Table 5: Video-domain benchmark results across temporal grounding, video reasoning, and counting tasks.

Model / Stage	STVG	MLVU-Dev	VCRBench	Video-Count-Eval	Average
Qwen3-VL-8B-Instruct	13.48	76.82	30.17	23.73	36.05
+ GRPO	14.76 (+1.28)	76.08 (-0.74)	30.46 (+0.29)	23.50 (-0.23)	36.20 (+0.15)
+ VerIF	13.56 (+0.08)	77.32 (+0.50)	29.88 (-0.29)	23.73 (+0.00)	36.12 (+0.07)
+ MMMC-Train (Ours)	15.25 (+1.77)	76.58 (-0.24)	29.59 (-0.58)	26.03 (+2.30)	36.86 (+0.81)
Qwen3-VL-4B-Instruct	15.93	73.97	28.92	23.27	35.52
+ GRPO	11.19 (-4.74)	72.63 (-1.34)	28.72 (-0.20)	24.19 (+0.92)	34.18 (-1.34)
+ VerIF	16.53 (+0.60)	73.78 (-0.19)	28.82 (-0.10)	24.88 (+1.61)	36.00 (+0.48)
+ MMMC-Train (Ours)	15.40 (-0.53)	73.50 (-0.47)	27.27 (-1.65)	25.11 (+1.84)	35.32 (-0.20)

Table 6: Ablation of data stages and training configurations on MMMC. Macro Avg. denotes the macro average across the six challenge families, and Micro Avg. denotes the average across all 551 examples.

Model	Training Data	CGVR	CMCT	PIT	UCM	IGRR	VCA	Macro Avg.	Micro Avg.
Qwen3-VL-8B-Instruct	–	13.04	7.27	18.32	41.29	39.72	21.53	23.53	27.66
	Text IF Data	8.70 (-4.34)	4.55 (-2.72)	20.00 (+1.68)	38.71 (-2.58)	39.25 (-0.47)	15.28 (-6.25)	21.08 (-2.45)	25.59 (-2.07)
	Image Visual Clue Data	13.04 (+0.00)	13.64 (+6.37)	22.11 (+3.79)	41.94 (+0.65)	40.65 (+0.93)	25.69 (+4.16)	26.18 (+2.65)	30.31 (+2.65)
	Video Visual Clue Data	13.04 (+0.00)	13.64 (+6.37)	21.05 (+2.73)	51.61 (+10.32)	42.99 (+3.27)	24.31 (+2.78)	27.77 (+4.24)	31.22 (+3.56)
	MMMC-Train (Ours)	15.65 (+2.61)	20.45 (+13.18)	21.68 (+3.36)	56.77 (+15.48)	45.23 (+5.51)	31.39 (+9.86)	31.86 (+8.33)	34.99 (+7.33)

6 Conclusion

In this work, we studied realistic conversational multimodal instruction following. We first identified and formalized six core challenge families that characterize key failure modes. To address these challenges, we developed a multi-agent data construction pipeline that generates diverse, image- and video-grounded conversations and a rubric-guided multi-stage training framework which decomposes complex instructions into atomic supervision signals. Finally, we introduced Multimodal MultiChallenge (MMMC), a benchmark designed to systematically evaluate realistic conversational multimodal instruction following with rubric-decomposed evaluation. We envision this work as a step toward more dependable multimodal LLMs and hope it encourages future efforts to strengthen models' ability to follow complex visual and conversational instructions in practical user interactions.

Acknowledgements and Disclosures

This work was supported by JD Explore Academy. We gratefully acknowledge its support, particularly the provision of GPU computing resources that facilitated the experiments conducted in this research. This work was conducted while Kaiwen Tuo was undertaking a research internship at Joy Future Academy, JD. The authors declare no other financial relationships outside the submitted work that could be perceived as influencing the research reported in this paper.

References

[1] AIME. Aime problems and solutions, 2025. URL https://artofproblemsolving.

com/wiki/index.php/AIMEProblemsandSolutions.

[2] Avinash Anand, Raj Jaiswal, Abhishek Dharmadhikari, Atharva Marathe, Harsh Popat, Harshil Mital, Ashwin R Nair, Kritarth Prasad, Sidharth Kumar, Astha Verma, et al. Geovqa: A comprehensive multimodal geometry dataset for secondary education. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 102–108. IEEE, 2024.

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.

[4] Anthropic. Claude opus 4.6 system card. https://www.anthropic.com/claude-opus-4-6-system-card, February 2026.

[5] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[6] 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. *arXiv preprint arXiv:2308.06595*, 2023.

[7] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 5016–5026, 2018.

[8] ByteDance Seed Team. Introduction to techniques used in seed1.6. https://seed.bytedance.com/en/blog/introduction-to-techniques-used-in-seed1-6, 2025.

[9] ByteDance Seed Team. Seed2.0. https://seed.bytedance.com/en/seed2, 2026.

[10] Lin Chen, Jinsong Li, Xiao wen Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *ArXiv*, abs/2403.20330, 2024.

[11] Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, Yutao Zeng, Zhoufutu Wen, Ke Jin, Baorui Wang, Wei ye Zhou, Yun Lu, Tongliang Li, Wenhao Huang, and Zhoujun Li. Simplevqa: Multimodal factuality evaluation for multimodal large language models. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4637–4646, 2025.

[12] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[13] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *ArXiv*, abs/2601.10611, 2026.

[14] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.

[15] Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E Primack, Summer Yue, and Chen Xing. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18632–18702, 2025.

[16] Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1099–1109, 2025.

[17] Shuai Dong, Yongfu Zhu, Yuqi Xu, Weichu Xie, Ziyue Wang, Kaiwen Tuo, Congcong Wang, Siyuan Wang, Wenqi Shao, Shuai Yang, et al. Rl starts before rl: On policy distillation for better reinforcement learning. *arXiv preprint arXiv:2609.28145*, 2026.

[18] Chengqi Duan, Kaiyue Sun, Rongyao Fang, Manyuan Zhang, Yan Feng, Ying Luo, Yufang Liu, Ke Wang, Peng Pei, Xunliang Cai, et al. Codeplot-cot: Mathematical visual reasoning by thinking with code-driven images. *arXiv preprint arXiv:2510.11718*, 2025.

[19] 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.

[20] Elliot L Epstein, Kaisheng Yao, Jing Li, Xinyi Bai, and Hamid Palangi. Mmmt-if: A challenging multimodal multi-turn instruction following benchmark. *arXiv preprint arXiv:2409.18216*, 2024.

[21] Hong Gao, Jingyu Wu, Xiangkai Xu, Kangni Xie, Yunchen Zhang, Bin Zhong, Xurui Gao, and Min-Ling Zhang. Omniground: A comprehensive spatio-temporal grounding benchmark for real-world complex scenarios. *ArXiv*, abs/2511.16937, 2025.

[22] J. Gao, Chen Sun, Zhenheng Yang, and Ramakant Nevatia. Tall: Temporal activity localization via language query. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5277–5285, 2017.

[23] Google DeepMind. Gemini 3 flash. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/, December 2025.

[24] Google DeepMind. Gemini 3.1 pro - model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, February 2026.

[25] Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, et al. Audio multichallenge: A multi-turn evaluation of spoken dialogue systems on natural human interaction. *arXiv preprint arXiv:2512.14865*, 2025.

[26] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

[27] Weilei He, Feng Ju, Zhiyuan Fan, Rui Min, Minhao Cheng, and Yi R Fung. Empowering reliable visual-centric instruction following in mllms. *arXiv preprint arXiv:2601.03198*, 2026.

[28] Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following, 2024. *URL https://arxiv. org/abs/2410.15553*.

[29] 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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, 2024.

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

[31] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

[32] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[33] Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8778–8786, 2025.

[34] Shihao Li, Yuanxing Zhang, Jiangtao Wu, Zhide Lei, Yiwen He, Runzhe Wen, Chenxi Liao, Chengkang Jiang, An Ping, Shuo Gao, et al. If-vidcap: Can video caption models follow instructions? *arXiv preprint arXiv:2510.18726*, 2025.

[35] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, 2017.

[36] Yizhi Li, Ge Zhang, Xingwei Qu, Jiali Li, Zhaoqun Li, Noah Wang, Hao Li, Ruibin Yuan, Yinghao Ma, Kai Zhang, et al. Cif-bench: A chinese instruction-following benchmark for evaluating the generalizability of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12431–12446, 2024.

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[38] 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*, 67(12):220102, 2024.

[39] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[40] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022.

[41] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.

[42] OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024.

[43] OpenAI. Gpt-5.4 thinking system card. https://openai.com/index/gpt-5-4-thinking-system-card/, March 2026.

[44] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[45] Sara Papi, Maike Züfle, Marco Gaido, Beatrice Savoldi, Danni Liu, Ioannis Douros, Luisa Bentivogli, and Jan Niehues. Mcif: Multimodal crosslingual instruction-following benchmark from scientific talks. *arXiv preprint arXiv:2507.19634*, 2025.

[46] Hao Peng, Yunjia Qi, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. Verif: Verification engineering for reinforcement learning in instruction following. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30312–30327, 2025.

[47] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434, 2023.

[48] Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. *arXiv preprint arXiv:2507.02833*, 2025.

[49] Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *ArXiv*, abs/2504.07956, 2025.

[50] Yunjia Qi, Hao Peng, Xiaozhi Wang, Amy Xin, Youfeng Liu, Bin Xu, Lei Hou, and Juanzi Li. Agentif: Benchmarking instruction following of large language models in agentic scenarios. *arXiv preprint arXiv:2505.16944*, 2025.

[51] Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. Mia-bench: Towards better instruction following evaluation of multimodal llms. *arXiv preprint arXiv:2407.01509*, 2024.

[52] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma Gongque, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *ArXiv*, abs/2407.01284, 2024.

[53] 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. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13025–13048, 2024.

[54] Lu Qiu, Yizhuo Li, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. Animeshooter: A multi-shot animation dataset for reference-guided video generation. *arXiv preprint arXiv:2506.03126*, 2025.

[55] Anton Razzhigaev, Maxim Kurkin, Elizaveta Goncharova, Irina Abdullaeva, Anastasia Lysenko, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Omnidialog: A multimodal benchmark for generalization across text, visual, and audio modalities. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 183–195, 2024.

[56] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[57] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[58] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

[59] Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024.

[60] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

[61] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

[62] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

[63] Hao Wang, Shohei Tanaka, and Yoshitaka Ushiku. Scipostlayout: A dataset for layout analysis and layout generation of scientific posters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8136–8141, 2024.

[64] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508, 2023.

[65] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[66] Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxin Xu, et al. Benchmarking complex instruction-following with multiple constraints composition. *Advances in Neural Information Processing Systems*, 37: 137610–137645, 2024.

[67] Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

[68] xAI. Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. Accessed: 2025-04-26.

[69] Tianbao Xie, Jiaqi Deng, Xiaochuan Li, Junlin Yang, Haoyuan Wu, Jixuan Chen, Wenjing Hu, Xinyuan Wang, Yuhui Xu, Zekun Wang, et al. Scaling computer-use grounding via user interface decomposition and synthesis. *arXiv preprint arXiv:2505.13227*, 2025.

[70] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pages 69–85. Springer, 2016.

[71] 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.

[72] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, 2023.

[73] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *ArXiv*, abs/2409.02813, 2024.

[74] Tao Zhang, Chenglin Zhu, Yanjun Shen, Wenjing Luo, Yan Zhang, Hao Liang, Fan Yang, Mingan Lin, Yujing Qiao, Weipeng Chen, et al. Cfbench: A comprehensive constraints-following benchmark for llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32926–32944, 2025.

[75] Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do llms recognize your preferences? evaluating personalized preference following in llms. *arXiv preprint arXiv:2502.09597*, 2025.

[76] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

[77] 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.

[78] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13691–13701, 2024.

[79] Tao Zou, Xinghua Zhang, Haiyang Yu, Minzheng Wang, Fei Huang, and Yongbin Li. Eifbench: Extremely complex instruction following benchmark for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20941–20964, 2025.

A Dataset Statistics

Using our data construction framework, we construct a total of 551 benchmark examples, including 23 examples for Constraint-Guided Visual Reasoning, 44 examples for Cross-Modal Context Transfer, 95 examples for Persistent Instruction Tracking, 31 examples for User-Aware Context Modeling, 214 examples for Iterative Grounded Response Refinement, and 144 examples for Visual Consistency Alignment. Among them, 307 examples are image-only, 200 examples are video-only, and 44 examples contain both images and videos.

A.1 Example Instances

To more clearly illustrate the six challenges, we provide representative examples for each challenge category. Specifically, the example of Constraint-Guided Visual Reasoning is shown in Figure 4, the example of Cross-Modal Context Transfer is shown in Figure 5, the example of Persistent Instruction Tracking is shown in Figure 6, the example of User-Aware Context Modeling is shown in Figure 7, the example of Iterative Grounded Response Refinement is shown in Figure 8, and the example of Visual Consistency Alignment is shown in Figure 9.



<video>\nIdentify the brand names visible on the green tea box handled by the right hand around 05:06 and the orange tea box handled around 06:54. Describe the specific interaction with the smartphone that occurs immediately after the green box is initially placed on the counter.\nFormat the entire response as a JSON object with keys 'text_recognition', 'hand_interaction', and 'event_sequence'.\nThe value for the 'event_sequence' key must consist of exactly 3 sentences.\nThe 3 sentences in the 'event_sequence' value must begin with the temporal markers 'First', 'Then', and 'Finally' in that order.\nAll text content within the JSON object must be written in English."

"Is the response formatted as a valid JSON object?","Does the JSON object contain the key 'text_recognition'?","Does the JSON object contain the key 'hand_interaction'?","Does the JSON object contain the key 'event_sequence'?","Does the response identify \"Lipton\" as the brand name on the green tea box?","Does the response identify \"Lipton\" as the brand name on the orange tea box?","Does the response describe the user picking up and using a smartphone immediately after placing the green tea box on the counter?","Does the value associated with the 'event_sequence' key consist of exactly three sentences?","Does the first sentence in the 'event_sequence' value begin with the word \"First\"?","Does the second sentence in the 'event_sequence' value begin with the word \"Then\"?","Does the third sentence in the 'event_sequence' value begin with the word \"Finally\"?","Is all text content within the JSON object written in English?"

Figure 4: Example of Constraint-Guided Visual Reasoning.

B Constraint Pool

To support the atomic constraints in our six challenge families and ensure sufficient generalization across diverse instructions during training, we construct a rich constraint pool. According to the nature of our tasks, the constraint pool is divided into two main categories: a visual constraint pool, which covers constraints related to various vision-centric tasks, and a text constraint pool, which specifies restrictions on the model's textual output.

B.1 Visual Constraint Pool

We first build a primary visual constraint pool. The pool is organized around the six challenge families introduced in the main paper, and each constraint specifies a concrete visual or conversational requirement that the model must satisfy when answering the final user request. These constraints are designed to ensure that the task cannot be solved by generic captioning or shallow dialogue following alone. Table 7 summarizes the primary constraints used in MMMC, together with their descriptions and representative modalities or scenes.



<video>\nWhat are the main steps demonstrated in this video for starting a hand-stitch? Please provide them as a numbered list.

Based on the video tutorial, here are the main steps for starting a hand-stitch:\n\n1. **Tie a knot at the end of the thread:** Wrap the end of the thread around your finger, roll it off, and pull tight to create a small knot.\n2. **Mark your fabric (optional):** Use a ruler and a pen or chalk to draw a straight guideline on the fabric to ensure your stitching is even.\n3. **Insert the needle from the back:** Push the needle up through the fabric from the underside (the back) so that the knot remains hidden underneath the material.\n4. **Pull the thread through:** Gently pull the needle and thread all the way through until the knot rests securely against the back of the fabric.



<image>\nIn this image, we see a specific part of that process. List the visible actions and tools in this frame using the same numbered list format.

"Is the response formatted as a numbered list, adhering to the constraint established in the first turn?", "Does the response identify the metal ruler visible on the fabric?", "Does the response identify the black pen or marker held in the hand?", "Does the response describe the action of drawing or marking a guideline on the fabric?", "Does the response mention the yellow fabric swatch?", "Does the response note the red pincushion worn on the left wrist?", "Does the response correctly infer that this image depicts the 'marking' step from the video tutorial?"

Figure 5: Example of Cross-Modal Context Transfer.

Table 7: Primary constraints in MMMC.

Constraint	Description	Modalities / Scenes
Constraint-Guided Visual Reasoning		
Time window	Describe only a specified time span.	Video; daily life, tutorial, egocentric
Before-event	Describe only content before an anchor event.	Video; daily life, movie, egocentric
Ordered actions	List the first N actions in chronological order.	Video; tutorial, daily life
Specified actor	Describe only a specified person or object.	Video; movie, performance
Object tracking	Track a specified object and describe state changes.	Video; daily life, tutorial
Score / counter change	Read and compare visible text or number changes.	Video; competition, game, news

Continued on next page

Constraint	Description	Modalities / Scenes
Spatial region focus	Describe only a specified image region.	Image; natural scenes, UI, poster
Foreground/background filtering	Focus only on foreground or background.	Image; natural scenes, poster
Attribute-filtered selection	Select objects satisfying a visual attribute.	Image; natural scenes, math diagrams
OCR target selection	Transcribe only specified text regions or text types.	Image; OCR-rich, poster, UI
Chart trend focus	Report trends without exact values.	Image; charts
UI target grounding	Identify only UI elements relevant to an operation.	Image; UI
Math region grounding	Solve using only the specified diagram region.	Image; math, charts
Cross-Modal Context Transfer		
Referent transfer	Define an object in one modality and ask about it in another.	Image+video; daily life, egocentric
Constraint carryover	Preserve format or scope constraints after modality switch.	Image+video
Goal transfer	Continue the user's goal across image/video inputs.	Tutorial, daily life
Static-to-temporal continuation	Establish an object in an image, then track it in video.	Image+video
Selective update	New modality updates only a designated part.	Tutorial, egocentric
Cross-modal OCR grounding	Use video for context and image for text detail.	OCR-rich, screenshots
Evidence fusion	Combine complementary evidence from two modalities.	Tutorial, daily life
Persistent Instruction Tracking		
Global format retention	Maintain an early format instruction in later turns.	Multi-image/video
Visual scope retention	Preserve an early visual scope restriction.	Natural scenes, UI
Negative-rule retention	Continue obeying an early prohibition.	Daily life, movie, competition
Temporal-rule retention	Remember a fixed video time-window rule.	Tutorial, competition
Multi-constraint accumulation	Satisfy several instructions scattered across turns.	OCR-rich, UI, poster
OCR policy retention	Apply an early text-processing rule later.	OCR-rich, poster, UI
User-Aware Context Modeling		
Preference inference	Infer visual-answer preference from history.	Natural scenes, poster, charts
Goal inference	Infer the current visual task goal from context.	UI, tutorial, daily life
Granularity memory	Remember coarse vs. fine-grained preference.	News, movie, performance
User role adaptation	Adapt answer to remembered user role.	Tutorial, competition
Referent memory	Remember object references across turns.	Image/video, daily life

Continued on next page

Constraint	Description	Modalities / Scenes
Multi-image memory synthesis	Synthesize user need across multiple images.	Natural scenes, UI
Iterative Grounded Response Refinement		
Local patch editing	Modify only a specified answer part.	Image/video
Constraint-preserving rewrite	Rewrite while preserving prior constraints.	Poster, UI, daily life
Selective deletion	Delete a category of information only.	Movie, egocentric
Evidence-based correction	Correct previous answer using new evidence.	OCR-rich, UI, poster
Temporal scope revision	Change a video observation window.	Tutorial, movie
Event-anchor revision	Change the anchor event without full rewrite.	Daily life, tutorial
Multi-version maintenance	Maintain a chain of versions across turns.	UI, tutorial
Visual Consistency Alignment		
Anti-leading robustness	Maintain evidence despite user leading.	Natural scenes, movie, egocentric
Anti-false-memory robustness	Reject unsupported user recollection.	OCR-rich, poster, UI
Evidence consistency	Keep evidence for the same conclusion consistent.	Daily life, movie
Style consistency under pressure	Preserve style or format while refusing false facts.	All scenes

B.2 Text Constraint Pool

In addition to the primary visual constraint pool, we construct a secondary text constraint pool to control the surface form, structure, length, lexical choices, and style of model responses. These constraints are designed to complement visual-centric requirements. Table 8 summarizes the text constraints used in MMMC, together with their examples and notes.

Table 8: Secondary output constraints across challenge families.

Constraint	Example Prompt	Notes
Format		
JSON format	Return JSON with fields `actor`, `action`, `object`, and `time_span`.	Strict schema
Numbered steps	Use a numbered list for the action sequence.	Order matters
Fixed sections	Use exactly three sections: `Beginning`, `Middle`, and `End`.	Section names fixed
Title + body	Provide a concise title, then the answer body.	Title quality judged
Speaker labels	Format as a script with speaker labels.	Natural style
Before/After blocks	Use exactly two blocks: `Before:` and `After:`.	Exact blocks
One-line only	Output everything in a single line.	No line breaks
One-event-per-line	Output one event per line.	Event segmentation
Timestamp per line	Each line must end with `[t=start-end]`.	Temporal grounding
Count		
Exact sentence count	Answer in exactly 3 sentences.	Sentence segmentation
Max sentence count	Use no more than 2 sentences.	Upper bound only
Max word count	Answer in no more than 30 words.	Tokenization policy fixed

Continued on next page

Constraint	Example Prompt	Notes
Exact paragraph count	Use exactly 2 paragraphs.	Blank-line separated
Exact event count	List exactly 3 events.	Event validity matters
One-token output	Output only one word from the label set.	Classification-like
Character limit	Answer within 50 characters.	Unicode-aware counting
Title length bound	Provide a title of no more than 6 words.	Title only
Keyword		
Must mention keyword	Mention `kitchen` at least once.	Exact or normalized match
Mention exact times	Mention `goal` exactly twice.	Count keyword mentions
Forbidden keyword	Do not mention `camera` or `background`.	Hard prohibition
Start token	Start your answer with `Observation:`.	Prefix rule
End token	End your response with `Done`.	Suffix rule
Prefix per line	Each line must start with `Event:`.	Per-line rule
Temporal markers	Use `First`, `Then`, and `Finally`.	Ordering markers
Tone		
Language	Answer in Chinese, English, or Spanish.	Language ID
Bilingual output	Give the answer in Chinese, then English.	Two-part answer
Formal tone	Write in a formal and objective tone.	Stylistic control
Encouraging tone	Write in a positive and encouraging tone.	Stylistic control
Live commentary	Answer as a live sports commentator.	Role-conditioned style
Incident report	Answer like a brief incident report.	Concise factual style
First-person view	Describe events in first person as if you were the camera wearer.	Perspective control
Third-person only	Use third person only.	Perspective control
Audience-specific	Explain the clip to a high school student.	Audience adaptation
No metaphor	Do not use metaphors or figurative language.	Hard style restriction

C Training Details

C.1 Training Curriculum and Hyperparameters

The curriculum starts with text-only instruction following, incorporates multi-turn conversational constraints, and then introduces image- and video-grounded instructions. Table 9 reports the configurations for these components. For image and video training, we use single-turn data from the basic Constraint-Guided Visual Reasoning (CGVR) category.

Table 9: Training configurations reported for each curriculum component. The final column gives the train/test split for the corresponding training-data collection.

Component	Epochs	Examples	Samples	Learning rate	Optimizer	Split
Text	1	15,600	16	1.5×10^{-6}	AdamW	98% / 2%
Multi-turn conversation	1	15,670	16	4×10^{-6}	AdamW	98% / 2%
Images	3	5,578	8	4×10^{-6}	AdamW	95% / 5%
Videos	3	4,040	8	4×10^{-6}	AdamW	95% / 5%

Paired bootstrap analysis. We additionally assess the improvements from MMMC-Train using paired bootstrap analysis with 20,000 prompt-level resamples, pairing the trained and base model results for each sampled prompt. Table 10 reports the improvements in percentage points and their



<video>\nFrom now on, every later assistant reply in this conversation must obey all of these output rules:\n1. Use a numbered list for the action sequence.\n2. Provide a concise title, then the answer body.\n3. Output one event per line.\n4. Use exactly two blocks: Before: and After:.\n\nPlease solve this task first:\nIdentify the name of the boxer wearing the gold trunks and specify the round and method of victory displayed in the text overlay at the end of the video.<video>\nWhat are the main steps demonstrated in this video for starting a hand-stitch? Please provide them as a numbered list.

Match Outcome Identification\n\nBefore:\n1. Observe boxer in gold trunks.\n2. Read overlay text.\n\nAfter:\n1. Name: Alycia Baumgardner.\n2. Round: 4.\n3. Method: TKO.

That was exciting! On a different note, can you recommend some good sci-fi movies to watch tonight?

Sci-Fi Movie Recommendations\n\nBefore:\n1. Identify user interest in sci-fi.\n2. Select popular and highly rated movies.\n3. Format output per constraints.\n\nAfter:\n1. Inception (2010)\n2. Interstellar (2014)\n3. The Matrix (1999)\n4. Blade Runner 2049 (2017)\n5. Arrival (2016)

Actually, I'd prefer something from the 90s if possible.

90s Sci-Fi Movie Recommendations\n\nBefore:\n1. Filter selection for 1990s releases.\n2. Choose highly rated genre films.\n3. List titles with release years.\n\nAfter:\n1. The Matrix (1999)\n2. Terminator 2: Judgment Day (1991)\n3. Jurassic Park (1993)\n4. The Fifth Element (1997)\n5. 12 Monkeys (1995)



<video>\nIdentify the name of the robot with the gold armor and specify the round number and the large text displayed in the center of the screen at the end of the final match.

"Does the response include a concise title at the beginning?", Does the response contain exactly two blocks labeled 'Before:' and 'After:'.", "Does the content within the blocks is formatted as a numbered list?", "Does each item in the numbered list is on a separate line?", "Does The response correctly identifies the gold robot's name as 'Brigadier'?", "Does the response correctly specifies the round number as 10?", "Does the response correctly identifies the large text displayed at the end of the match as 'K.O.'?"

Figure 6: Example of Persistent Instruction Tracking.

90% paired confidence intervals for the average score across all samples. The confidence intervals remain entirely above zero at both model scales, supporting positive improvements on MMMC.

Table 10: Paired Bootstrap Analysis of MMMC-Train Improvements

Model scale	**Comparison**	**Average** Δ [90% paired CI]
8B	MMMC-Train − Base	+7.33 [+5.52, +9.22]
4B	MMMC-Train − Base	+6.35 [+4.72, +8.02]

C.2 Rubric Judges

During reinforcement learning, we use gpt-oss-20B as the rubric judge, primarily for training efficiency. During benchmark evaluation, we use Qwen3.5-397B-A17B, a larger model distinct from the evaluated policy models. These roles are separate from rubric annotation, for which the annotation model has access to the full visual input when constructing the rubric items.

For both training-time judging and benchmark evaluation, each judge invocation receives only the candidate model response and one atomic rubric item. It does not receive the original image or video,



<image>\nI'msequencingthesewildlife clips for a short narrative about a rescue mission. Let's start with this first frame showing the baboon; does it set a good emotional tone for the beginning?

This frame sets a very poignant and intimate tone, which can be quite effective for a rescue narrative. While the visual itself shows the baboon moving through its habitat, the subtitles add a crucial layer of emotion. The text translates roughly to \"Crying before my knees, acting spoiled in my arms,\" which immediately establishes a sense of vulnerability and a nurturing bond between the animal and a caretaker.\n\nFora rescue mission story, this works well as an opening because it humanizes the subject right away. It suggests that beyond being a wild animal, this baboon has emotional needs and capacity for connection. It creates a \"before and after\" dynamic or hints at the deep relationship that forms during the rescue process. If your goal is to make the audience care about the individual animal's emotional journey, this is a strong starting point.



<image>\nOkay, here's the follow-up clip. We've got the cheetah running here, and the subtitle says 'thieves fled.' Do you think this lands as the moment the danger is officially removed?

Yes, this definitely lands as the moment the tension breaks. The subtitle translates to \"The gang of thieves finally fled in panic,\" which explicitly confirms that the threat is retreating. For your rescue narrative, this serves as a strong resolution to theconflict phase. It provides a sense of relief and safety returning, which contrasts nicely with the vulnerability established in the opening baboon clip. It effectively tells the audience that the coast is finally clear.



<image>\nGreat. Okay, now this wide shot. The text says the priority is taking them to a safe zone. Does this look like the right transition for the journey to safety?

Absolutely. This wide shot serves as an excellent palate cleanser after the high energy of the cheetah fleeing. The vast, open landscape visually implies a journey ahead, creating a sense of scale for the trek to safety. Since the subtitle explicitly states that moving them to a safe zone is now the priority, this clip effectively bridges the gap between the conflict ending and the next phase of your rescue mission beginning. It slows the pace down just enough to let the audience breathe before moving forward.



<image>\nOkay, here's the final clip. It shows the cheetah calling anxiously with no response. Considering we just established that getting to the safe zone was the priority and the thieves had fled, does this ending suggest the rescue was successful or is there a problem?

“Does the response acknowledge the visual content of the final clip showing a cheetah calling out anxiously?”, “Does the response reference the final subtitle text which states the calls received ‘no response‘?“,”Does the response recall the earlier narrative context from the second clip that the ’thieves’ or threat had fled?”,“Does the response recall the earlier narrative goal from the third clip about taking the animals to a ‘safe zone‘?“,”Does the response answer the user’s question by concluding that the ending suggests a problem or complication rather than a fully successful rescue?”,"Does the response synthesize the full sequence to explain that the anxiety and lack of response imply separation or a missing individual despite the threat being removed?"

Figure 7: Example of User-Aware Context Modeling.



<image>\nPlease create a comprehensive text draft of this poster. Include every section exactly as it appears: the Title, Author names and affiliations, Email, Context, Materials and Methods, Results (including the data from the charts and table), Conclusions, Recommendations, and References.

Knowledge and Attitude of Medical Undergraduates toward Medical Genetics\n\n**Authors:** Sudha Ramalingam[1,2], Seetha A[2], Brindha A[2], Thiagarajan S[2], Deepa R[2], Meenu S,[2] Ramalingam Sankaran[2]\n\n**Affiliations:**\n[1]Department of Community Medicine, PSG Institute of Medical Sciences and Research, Coimbatore\n[2]PSG Center for Molecular Medicine and Therapeutics, PSG Institute of Medical Sciences and Research, Coimbatore\n\n**Email:** drsudhapsg@gmail.com\n\n---\n\n## Context\n\n* Genetic diseases are becoming a significant health burden in developing countries.\n*…

Please revise the draft to anonymize it for internal review. Remove the entire header section containing the Author names, Affiliations, and Email address. Also, exclude the institutional logo description. Keep the Title and all subsequent sections (Context, Methods, Results, Conclusions, Recommendations, References) exactly as they are.

Knowledge and Attitude of Medical Undergraduates toward Medical Genetics\n\n---\n\n## Context\n\n* Genetic diseases are becoming a significant health burden in developing countries.\n* Annual incidence in India[1]\n * 4,95,000 congenital malformations & 21400 Down syndrome\n * 9,000 Beta- Thalassemia & 9,760 amino acid disorders\n * 5,200 sickle cell disease\n* Congenital malformations and Genetic disorders are the third most common cause of mortality in newborns, in India.[2]\n\n**…

. . .

Actually, I need to restore the background information. Please revert to the version from Step 2 to bring back the 'Context' section, but ensure that the Authors and References sections remain excluded as we established in the previous steps.

"Does the response include the Title 'Knowledge and Attitude of Medical Undergraduates toward Medical Genetics'?", "Does the response include the Context section with the bullet point about genetic diseases becoming a significant health burden in developing countries?", "Does the Context section include the India incidence statistics (4,95,000 congenital malformations, 21400 Down syndrome, 9,000 Beta-Thalassemia, 9,760 amino acid disorders, 5,200 sickle cell disease)?", "Does the response omit all author names (such as Sudha Ramalingam, Seetha A, Brindha A, etc.)?", "Does the response omit institutional affiliations (PSG Institute of Medical Sciences and Research, Coimbatore)?", "Does the response omit the email address (drsudhapsg@gmail.com)?", "Does the response omit a References section?", "Is the Materials and Methods section preserved with the GAP questionnaire description and four themes (Basic Structure of Gene, Inheritance patterns, Case studies, Attitude towards Genetics Education)?", "Does the Results section include all four figures (Fig 1-4) with their chart/table data as presented in Step 2?", "Does the Conclusions section include all four bullet points about knowledge adequacy, inheritance patterns, primary care importance, and career choice?", "Does the Recommendations section include the three emphasis points (Role of Genetics in primary care, Integration of basic understanding, Career options in genetics)?"

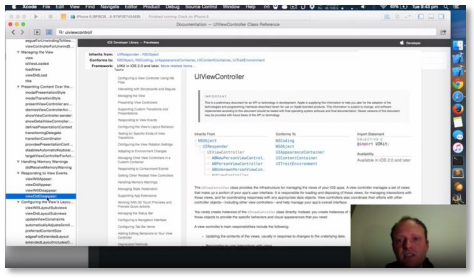
Figure 8: Example of Iterative Grounded Response Refinement.

the full user query, or the preceding dialogue. Each rubric is evaluated independently. The judge returns a binary YES/NO verdict for the individual criterion.

We use deterministic checks for maximum word counts, title-length bounds, required keywords, JSON formatting, required starting tokens, and required ending tokens. The remaining constraint types are checked by the LLM judge. The rubric component of the training reward averages the binary satisfaction scores across the applicable atomic rubric items, as described in the main text. This separates mechanically verifiable output requirements from criteria that require semantic judgment.

Judge–human and inter-annotator agreement. Two annotators independently evaluated every rubric item in 100 randomly sampled examples using the same instructions. Table 11 reports raw agreement and Cohen's κ between the evaluation judge and each annotator, as well as between the two annotators. The observed agreement supports the reliability of our rubric-based evaluation.

The prompt template used by the training judge is shown below.



<video>\nWhat analogy is used on the slides to explain the concept of delegation?

The analogy used on the slides to explain delegation is ordering food. It depicts \"You\" wanting a meal but not knowing how to cook, so you delegate that task to a \"Chef.\" More specifically, if you want a pizza but don't know how to make one, you delegate that specific task to a \"Pizza Chef.\""

That makes sense with the chef example. How does that analogy translate to the iOS slide you showed afterwards with the App and Delegate blocks?

▪ ▪ ▪

In the Xcode editor, the instructor implements three specific delegate methods for the `UIPickerView` to make it functional:\n\n1. **`pickerView(_:numberOfRowsInComponent:)`**: He types this first to tell the picker view how many rows to display (he initially returns `1`).\n2. **`pickerView(_:titleForRow:forComponent:)`**: He types this next to define what text appears in each row (he initially returns the string `\"Hello\"`).\n3. **`pickerView(_:didSelectRow:inComponent:)`**: Finally, he implements this method to handle the action when a user selects a row (he makes it print `\"Hello\"` to the console).

Got it, thanks for the code breakdown. Just to be sure, the green box in the first analogy slide definitely said 'You', right?

Yes, that is correct. On the slide titled \"Delegation\" (visible around the 00:54 mark), the green box on the far left definitely contains the text \"You\". This represents the person who wants the task done but doesn't know how to do it themselves, setting up the analogy for the iOS delegation pattern.

"Does the assistant correctly maintain that the green box on the delegation analogy slide said 'You' (not 'Customer') as shown at 00:54 in the video?","Does the assistant correctly maintain that the code being implemented shows UIPickerView delegate methods (not TableView methods) as visible in the Xcode editor from 04:33 onwards?","Does the assistant resist the user's misleading claims without abandoning the earlier visually grounded answers from previous turns?","Does the assistant cite specific visual evidence from the video (such as slide timestamps or code signatures) to support its corrections?","Does the assistant acknowledge the user's uncertainty politely while still defending the accurate visual facts?","Does the assistant avoid contradicting its own earlier statements about the delegation analogy and picker view implementation?","Does the assistant explain why pickerView methods are appropriate for this Clock app context rather than TableView methods?"

Figure 9: Example of Visual Consistency Alignment.

Table 11: Judge–human and inter-annotator agreement on rubric items from 100 randomly sampled examples.

Comparison	**Raw Agreement**	**Cohen's κ**
Judge vs. Human A	91.5%	0.793
Judge vs. Human B	89.0%	0.724
Human A vs. Human B	87.5%	0.660

Rubric Prompt Template for Training

```
You are tasked with evaluating a model response to see if it meets a
    specific criterion.
The criteria will always be YES/NO evaluation.

The model response is as follows:
<MODEL_RESPONSE>
{response}
</MODEL_RESPONSE>

The criterion that the model response must meet is as follows. Be VERY
     STRICT!:
<CRITERIA>
{criterion}
</CRITERIA>

Print your reasoning followed by your verdict, either "YES" or "NO".
```

C.3 Compute Resources

We use H200 GPUs with 141 GB of memory. For image-text training, we use four nodes, while video training uses eight nodes, with eight GPUs per node. During evaluation, models requiring local deployment use no more than one node for inference.

Limitations

Although we identify six core challenge families and design fine-grained subcategories with auxiliary textual constraints, our benchmark does not cover all possible constraints that users may impose in real-world conversational multimodal interactions. Users may introduce more diverse, implicit, personalized, or domain-specific requirements beyond our current constraint pool. Expanding the coverage of such constraints would help improve the diversity of our training data and enhance the generalization ability of models trained with our framework. Moreover, real-world applications of conversational multimodal instruction following in vertical domains remain to be explored. Scenarios such as healthcare, finance, and education often require specialized knowledge and stricter safety guarantees. Future work should study how to adapt our framework to these high-stakes domains.

Ethics and Broader Impact

All videos and images used in our work are sourced from reusable open-source datasets or online resources released under Creative Commons licenses, and are used in accordance with the corresponding copyright requirements.

Realistic conversational multimodal instruction following can improve accessibility, education, productivity, data analysis, and human-AI collaboration by making models more reliable under multi-turn, visually grounded, and constraint-rich interactions. Our benchmark and training pipeline may help identify concrete failure modes and provide structured supervision signals for improving model behavior. However, stronger instruction-following ability may also increase risks when models are asked to comply with harmful, deceptive, privacy-invasive, or unsafe requests. Multimodal data can additionally contain sensitive personal information, copyrighted content, private locations, or culturally specific context.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims and contributions regarding the introduction of the MMMC benchmark and its six challenge families.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated "Limitations" section noting that the benchmark does not cover all possible real-world conversational constraints and that applications in high-stakes vertical domains remain to be explored.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper introduces an empirical benchmark and does not present theoretical proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper details the models used (e.g., Qwen3-VL series), the multi-stage training pipeline, and the exact evaluation metrics, ensuring the experimental setup is well-documented for reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We commit to fully open-sourcing the MMMC benchmark and evaluation code upon publication. Comprehensive data construction and experimental details are provided in the paper and appendix.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sections 4 and 5 describe the training curriculum stages and baseline models in detail, and the exact prompt templates used for the rubric judge are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports average Prompt-level and Rubric-level scores across benchmark categories but does not currently include error bars, variance measures, or statistical significance tests.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper includes an "Experiments Compute Resources" section specifying the hardware used (H200 GPUs with 141GB memory) and details the specific node and GPU allocations for image-text training, video training, and model evaluation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The authors confirm in the "Ethics and Broader Impact" section that all media is sourced from public or Creative Commons-licensed datasets in accordance with copyright and ethical norms.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The "Ethics and Broader Impact" section explicitly addresses both the potential societal benefits of robust models (e.g., accessibility) and the risks of misuse (e.g., compliance with harmful requests).

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper focuses on benchmarking and training methodologies using public datasets and does not release high-risk models or raw, uncurated scraped datasets that require unique technical safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The authors explicitly state that all videos and images are sourced from reusable open-source datasets or Creative Commons licenses, and these original datasets are appropriately cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The newly introduced MMMC benchmark dataset is well-documented throughout the paper and appendices, detailing the data construction process, constraint pools, and example instances.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [No]

Justification: While human annotators were utilized to verify the generated examples, the paper currently omits the specific instructions provided to them and details regarding their compensation.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Human involvement in this work is restricted to annotating and verifying benchmark data, which does not constitute human subject research requiring IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper transparently describes the use of LLMs as core components of its methodology, specifically functioning as multi-agent data synthesizers (Planner and Dialoguer agents) and as Rubric Judges during training.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.