

# MM-JudgeBias: A Benchmark for Evaluating Compositional Biases in MLLM-as-a-Judge

Sua Lee<sup>1\*</sup>, Sanghee Park<sup>2,3</sup>, Jinbae Im<sup>2,3\*†</sup>

<sup>1</sup>Seoul National University, <sup>2</sup>NAVER Cloud AI, <sup>3</sup>KAIST AI

## Abstract

Multimodal Large Language Models (MLLMs) have been increasingly used as automatic evaluators—a paradigm known as *MLLM-as-a-Judge*. However, their reliability and vulnerabilities to biases remain underexplored. We find that many MLLM judges fail to reliably integrate key visual or textual cues, yielding unreliable evaluations when evidence is missing or mismatched, and exhibiting instability under semantically irrelevant perturbations. To address this, we systematically define *Compositional Bias* in MLLM-as-a-Judge systems and introduce **MM-JudgeBias**, a benchmark for evaluating it. MM-JudgeBias introduces controlled perturbations across Query, Image, and Response, and evaluates model behavior via two complementary metrics: *Bias-Deviation (BD)* for sensitivity and *Bias-Conformity (BC)* for stability. Our dataset of over 1,800 curated and refined multimodal samples, drawn from 29 source benchmarks, enables a fine-grained diagnosis of nine bias types across diverse tasks and domains. Experiments on 26 state-of-the-art MLLMs reveal systematic modality neglect and asymmetric evaluation tendencies, underscoring the need for more reliable judges.

[Project Page](#) | [Code](#) | [Dataset](#)

## 1 Introduction

Multimodal Large Language Models (MLLMs) (Caffagni et al., 2024; Achiam et al., 2023; Comanici et al., 2025; Bai et al., 2025b) have recently demonstrated remarkable capabilities in integrating vision and language, enabling complex perception and reasoning across modalities. Beyond serving as task solvers, MLLMs have increasingly been used as automatic judges—a paradigm known as *MLLM-as-a-Judge*—to assess multimodal generations such as captioning, visual question answering,

\*Equal contribution.

†Corresponding Author: [jinbae.im@navercorp.com](mailto:jinbae.im@navercorp.com).

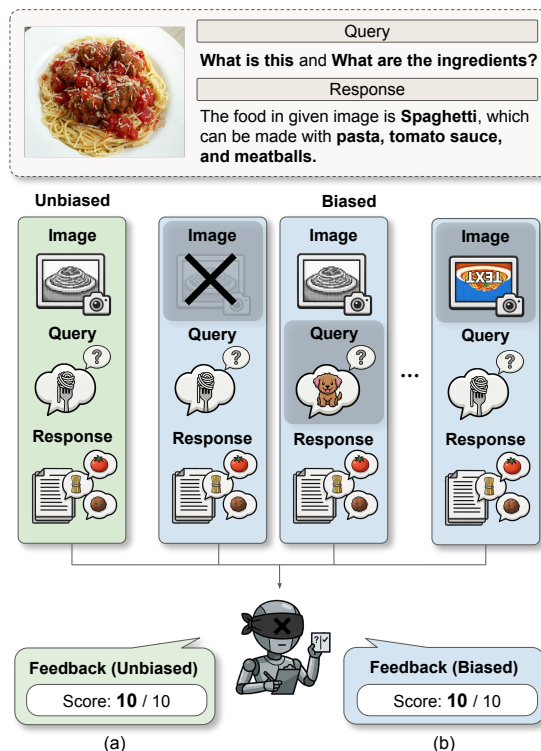


Figure 1: **Illustration of the Compositional Bias.** (a) Unbiased: A valid evaluation necessitates joint reasoning across all components (Image, Query, and Response); for example, answering “What is this?” inherently requires the judge to verify the response against the specific image provided. (b) Biased: Scenarios where essential grounding evidence such as the image or the original query is removed or replaced with irrelevant content. We observe that the *Compositional Bias* is prevalent in many MLLMs, leading them to assign a near-perfect score to biased scenarios despite the fundamental lack of context or semantic misalignment. (See Appendix G for actual examples.)

and visual reasoning (Chen et al., 2024a; Pu et al., 2025; Li et al., 2024b). While early studies (Zhang et al., 2023) have primarily focused on employing generalist MLLMs as general-purpose judges, recent efforts have shifted toward developing judge-specialized (critic) models (Lee et al., 2024; Xiong

et al., 2025), which are fine-tuned to provide feedback and judgments. However, despite substantial progress in benchmarking MLLMs in judging scenarios (Chen et al., 2024a; Pu et al., 2025; Zeng et al., 2025; Li et al., 2025; Ruan et al., 2025), the reliability and fundamental biases of MLLM-based judges remain underexplored, in contrast to the extensive studies on bias in LLM-as-a-Judge (Ye et al., 2025; Shi et al., 2025; Li et al., 2024b).

Recent studies have increasingly revealed that MLLMs often struggle to balance and integrate information across modalities, primarily exhibiting “modality biases” where models rely on spurious textual correlations to reach a correct answer (Chen et al., 2024c; Zhao et al., 2025; Zheng et al., 2026; Lin et al., 2025; Zheng et al., 2025). While these biases are typically viewed as limitations in task-solving accuracy, we argue that they evolve into a more insidious failure of verification integrity within the MLLM-as-a-Judge paradigm. Unlike a task-solver, the primary mandate of a judge is not mere probabilistic prediction, but the conditional verification of alignment between textual queries, visual evidence, and candidate responses.

We identify that judgment failures go beyond modality bias and stem from **Compositional Bias**: a systematic tendency where a judge fails to correctly integrate and reason over multiple components (query, image, and response), instead relying on partial, misaligned, or spurious compositions, as illustrated in Figure 1. Compositional bias does not merely reflect incorrect guessing by the judge; rather, it indicates situations where the judge actively endorses erroneous or ungrounded responses by failing to properly ground its judgment in the provided context, compromising the reliability of the entire evaluation pipeline (Xiong et al., 2026; Laskar et al., 2025). This motivates the need for a systematic benchmark that explicitly measures how reliably MLLM judges attend to and integrate visual, textual, and response information as a judge.

To address this, we introduce **MM-JudgeBias**, which stands for **MLLM-as-a-Judge Bias** Evaluation Benchmark, a comprehensive benchmark designed to measure the compositional biases in MLLM-as-a-Judge. We define a taxonomy of nine bias types across three critical dimensions: (1) *Integrity*: Does the judge consider all components holistically? (2) *Congruity*: Does the judge detect and appropriately address semantic contradictions between components? (3) *Robustness*: Does the judge maintain a stable judgment de-

spite semantic-preserving variations? Motivated by Ye et al. (2025), our framework systematically injects controlled bias perturbations and quantifies the model’s vulnerability to biases. Specifically, we utilize two complementary metrics: *Bias-Deviation (BD)* for sensitivity to disruptions in Integrality and Congruity, and *Bias-Conformity (BC)* for stability in Robustness.

To encompass a wide variety of tasks and domains, we structure our benchmark around two axes: four task types to assess different multimodal abilities, and 12 domain types to capture diverse visual contexts. We curate source data from 29 benchmarks and refine them for the MLLM-as-a-Judge setting, resulting in a benchmark set of 1,804 samples that cover diverse query tasks and difficulty levels. To comprehensively analyze biases in MLLM-as-a-Judge, we assess 26 MLLMs including closed-source, open-source, and critic models. Our evaluation results demonstrate that even state-of-the-art models struggle to jointly consider image inputs and text queries.

In summary, this work makes the following three key contributions. First, we demonstrate the emergence of the *Compositional Bias* that arises from the interaction between components of the judgment, such as text queries, image inputs, and responses. We also introduce *MM-JudgeBias*, a systematic benchmark that evaluates the compositional biases in MLLM-as-a-Judge through nine bias types and two complementary metrics: Bias-Deviation (BD) and Bias-Conformity (BC). Finally, we perform an extensive empirical analysis of 26 *state-of-the-art MLLMs*, uncovering that the compositional bias is a systemic issue pervasive even in the most advanced reasoning-heavy models.

## 2 Related Work

### 2.1 Bias and Reliability in LLM-as-a-Judge

The LLM-as-a-Judge framework has become central to automatic evaluation in generation tasks (Zheng et al., 2023; Gu et al., 2024; Li et al., 2024b). While early benchmarks showed high agreement with human judgments, subsequent analyses revealed systematic biases that undermine reliability. Recent studies demonstrate that LLM judges are influenced by superficial cues such as position, verbosity, and stylistic patterns, rather than actual content quality (Ye et al., 2025; Shi et al., 2025). Ye et al. (2025) identified 12 distinct bias types in large-scale evaluation; Shi et al. (2025) and Li

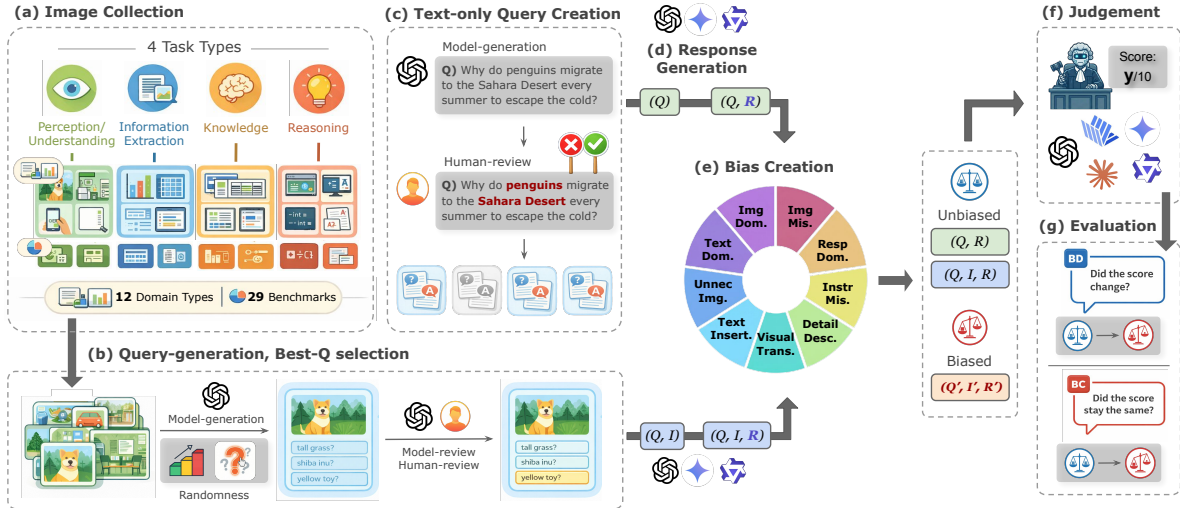


Figure 2: **Overview of the MM-JudgeBias construction and evaluation pipeline.** (a–b) We construct image sets from 29 source benchmarks covering four task types and 12 domain types, and then generate queries tailored to our bias evaluation setting through a comprehensive model-and-human verification framework to ensure high quality. (c) A text-only evaluation set is independently constructed via a parallel generation and verification process. (d–e) Based on nine defined bias types, we apply input perturbations to transform original unbiased samples—triplets  $(Q, I, R)$  or text-only pairs  $(Q, R)$ —into their corresponding biased variants  $(Q', I', R')$ . (f–g) Finally, we evaluate 26 MLLM judges using our proposed Bias-Deviation (BD) and Bias-Conformity (BC) metrics to systematically quantify the impact of Compositional Bias on their judgment reliability.

et al. (2026b) showed that prompt order and rubric phrasing significantly alter scores. Bias also arises from self-similarity between judge and generator models (preference leakage) and model-specific favoritism (Li et al., 2026a; Spiliopoulou et al., 2025). Cross-lingual and domain analyses confirm that such biases persist even under controlled conditions (Chen et al., 2024b; Gao et al., 2025). Collectively, these findings indicate that detecting and quantifying bias is crucial to maintain credibility in automated evaluation.

However, almost all prior studies remain limited to text-only scenarios. Existing approaches measure linguistic robustness but overlook multimodal evaluation, where visual and textual cues jointly determine correctness. This gap motivates the need to extend bias analysis into multimodal judging, where biases may manifest through modality dependence or neglect.

## 2.2 Bias and Reliability in MLLM-as-a-Judge

Recent work extends evaluation to MLLMs, aiming to assess their judgment capability and reliability on multimodal contexts (Chen et al., 2024a; Pu et al., 2025; Laskar et al., 2025; Zeng et al., 2025; Xiong et al., 2026; Li et al., 2025; Yasunaga et al., 2025; Ruan et al., 2025). However, biases in MLLM-as-a-Judge remain underexplored. Chen

et al. (2024a) analyzed egocentric, position, and length biases in MLLMs; however, such biases have been examined in LLM-as-a-Judge, leaving unique biases arising in multimodal contexts unaddressed. Hwang et al. (2025) defined several visual biases and evaluated the vulnerability of MLLM judges to the biases; however, their scope was limited to biases related to visual transformations in text-to-image evaluation. In this work, we introduce compositional biases in general multimodal judgment settings from the perspective of integrality, congruity, and robustness, and propose two metrics to systematically measure these biases. Further details on the positioning of our work are presented in Appendix M.

## 3 MM-JudgeBias: MLLM-as-a-Judge Bias Evaluation Benchmark

We introduce MM-JudgeBias, a benchmark built to quantify vulnerabilities to compositional biases in MLLM-as-a-Judge. MM-JudgeBias is designed around the central hypothesis that a reliable judge should attend to and integrate all available components—*Query, Image, and Response*—when producing its judgment. In this section, we introduce task formulation (3.1), evaluation metrics (3.2), bias taxonomy (3.3), and data construction

	Bias Type	Targeted Reliability (Purpose)	Perturbation Strategy	Metric
Integrity	<b>Text-Dominance</b>	Assesses over-reliance on linguistic cues when visual grounding is absent.	Replaces the image with a null image (black image).	BD
	<b>Image-Dominance</b>	Tests if the judge ignores the original query in favor of visual content.	Replaces the query with a null text (“”).	BD
	<b>Response-Dominance</b>	Evaluates if scores are assigned based on response fluency alone, ignoring all context.	Replaces both the query and image with null inputs (“” and a black image).	BD
Congruity	<b>Instruction-Misalignment</b>	Probes query-response conflict by checking for semantic mismatch between the query and response.	Replaces the query with a random, unrelated sample.	BD
	<b>Image-Misalignment</b>	Probes image-response conflict by checking for semantic mismatch between the image and response.	Replaces the image with a random, unrelated sample.	BD
Robustness	<b>Detail-Description</b>	Examines if providing redundant, aligned visual information inflates the score.	Appends a detailed caption of the given image to the query.	BC
	<b>Unnecessary-Image</b>	Tests the ability to recognize and ignore irrelevant visual inputs in text-only tasks.	Adds a random, unrelated image to a text-only task.	BC
	<b>Visual-Transformation</b>	Measures invariance to low-level visual transformations that preserve semantics.	Applies a diverse set of semantic-preserving augmentations to the image.	BC
	<b>Texture-Insertion</b>	Probes over-sensitivity to textual cues embedded directly within the visual modality.	Overlays query-related keywords or the query text itself onto the image.	BC

Table 1: **The MM-JudgeBias Taxonomy.** Biases are grouped into three dimensions: *Integrity* and *Congruity* evaluate the model’s sensitivity to semantic disruptions, while *Robustness* assesses judgment stability against semantic-preserving perturbations. (See Appendix F and D for detailed visual examples and perturbation strategies).

pipeline (3.4) of MM-JudgeBias.

### 3.1 Task Formulation

Let a judgment instance be a multimodal triplet  $(Q, I, R)$ , where  $Q$  denotes the text-based query (used interchangeably with question or instruction),  $I$  represents the associated visual context (image), and  $R$  is the model-generated response being evaluated. An MLLM-as-a-Judge model, denoted as  $f_\theta$ , serves as a scoring function:

$$f_\theta : (Q, I, R) \rightarrow y, \quad y \in \{1, 2, \dots, S\},$$

where  $y$  is a scalar score representing the quality of  $R$  given  $Q$  and  $I$ , and  $S$  is the maximum scale (e.g.,  $S = 10$ ). We adopt the 10-point scoring convention following recent open-ended evaluation protocols (Liu et al., 2023; Yu et al., 2024; Chen et al., 2024a). We provide detailed prompts in Appendix E.

To evaluate the susceptibility of  $f_\theta$  to biases, MM-JudgeBias pairs each unbiased triplet  $(Q, I, R)$  with a corresponding biased triplet  $(Q', I', R')$ . The biased triplet is generated by injecting a controlled perturbation into one or more components to evaluate whether the model’s judgments arise from a holistic integration of  $Q$ ,  $I$ , and

$R$ , rather than from biased or partial consideration of these factors.

The judge model thus produces two comparative scores: the reference score  $y = f_\theta(Q, I, R)$  and the perturbed score  $\hat{y} = f_\theta(Q', I', R')$ . We categorize the expected behavior of a reliable judge into two paradigms:

- 1. Sensitivity to Semantic Distortion (Integrity & Congruity):** When a perturbation disrupts the essential semantic alignment, such as modality removal (integrity) or inter-modal mismatching (congruity), a reliable judge should reflect this deficiency by decreasing the score ( $\hat{y} < y$ ).
- 2. Invariance to Semantic-Preserving Perturbations (Robustness):** When a perturbation introduces changes that do not affect the objective quality of the response such as minor visual transformations or redundant captions, the score should remain stable ( $\hat{y} \approx y$ ).

Based on these principles, we develop quantitative measures to evaluate the degree of score deviation and stability across different bias categories.

### 3.2 Evaluation Metrics

To effectively measure robustness under biased conditions, we propose two complementary metrics, *Bias-Deviation (BD)* and *Bias-Conformity (BC)*, which quantify the model’s behavioral response to semantic distortions and semantic-preserving perturbations, respectively.

**Bias-Deviation (BD).** The BD metric is designed to assess whether a judge responds appropriately to semantic distortions (e.g., modality removal or misalignment) by lowering its judgment score. For a dataset  $\mathcal{D}$  consisting of  $(y, \hat{y})$  pairs, BD is defined as follows:

$$\text{BD} = \mathbb{E}_{(y, \hat{y}) \sim \mathcal{D}} \left[ \frac{(y - \hat{y})_+}{y - 1} \mid y > 1 \right],$$

where the positive deviation between  $y$  and  $\hat{y}$  is normalized by the maximum positive deviation. Note that samples with  $y = 1$  are excluded from the BD computation, as no further score reduction is possible. A higher BD indicates that the model is sensitive to the injected bias and appropriately adjusts its judgment. For bias types associated with *Integrity* and *Congruity*, we use BD to measure the reliability of MLLM judges.

**Bias-Conformity (BC).** In contrast, the BC metric is devised to evaluate the stability of the model’s judgment against perturbations that should not affect the objective quality of the response (e.g., visual style changes or redundant captions). We define BC as the degree of score invariance:

$$\text{BC} = \mathbb{E}_{(y, \hat{y}) \sim \mathcal{D}} \left[ 1 - \frac{|y - \hat{y}|}{\max(y - 1, S - y)} \right],$$

where the absolute difference between  $y$  and  $\hat{y}$  is normalized by the upper bound on the score change. A high BC score signifies that the model is robust to irrelevant variations, maintaining a consistent evaluative standard. For bias types related to *Robustness*, an ideal evaluator should achieve a BC score near 1.0, indicating that its judgment is not swayed by superficial or non-essential modifications. As these perturbations are designed to be semantically invariant, the theoretical expectation for score deviation is minimal, inherently leading to a relatively high-value distribution of BC scores that characterizes a robust judge.

**Metric Selection and Interpretation.** It is important to note that BD and BC are not competing metrics but are applied selectively based

on the nature of the bias. We define a reliability score for a judge model  $f_\theta$  by aggregating BD across semantic-disruptive categories and BC across semantic-preserving categories. This dual-lens approach allows us to distinguish between a model that is simply “volatile” (high deviation in all cases) and one that is “discerning” (high deviation only when semantically justified).

### 3.3 Bias Taxonomy

To evaluate the multifaceted reliability of MLLM-as-a-Judge, we define nine bias types categorized into three functional dimensions based on the expected behavior of a reliable judge: *Integrity*, *Congruity*, and *Robustness*. Table 1 provides a detailed overview of the taxonomy, including the perturbation strategies and the metrics.

**Integrity-targeted Biases** test the model’s awareness of modality necessity—whether it recognizes that each input is strictly required for evaluation. We exclude modalities to probe modality dominance such as *Text-*, *Image-*, and *Response-Dominance*. A reliable judge is expected to penalize these information-deficient cases (high BD).

**Congruity-targeted Biases** probe cross-modal reasoning by checking whether the judge detects semantic contradictions and appropriately reflects them in its judgment. This involves cases where the query or image is mismatched with the response (*Instruction/Image-Misalignment*). Similar to *Integrity*, a logical judge should decrease its score when encountering such conflicts (high BD).

**Robustness-targeted Biases** assess the model’s stability against perturbations that do not alter the objective quality of a response. These involve redundant information (*Detail-Desc.*), irrelevant visual inputs (*Unnecessary-Img.*), or superficial variations (*Visual-Trans.*, *Texture-Insertion*). An ideal judge should maintain scoring stability regardless of these changes (high BC).

### 3.4 Data Construction

The construction of *MM-JudgeBias* follows a systematic pipeline designed to ensure task diversity, multimodal dependency, and rigorous quality control, as shown in Figure 2. We consolidate the process into four primary phases, yielding a final dataset of 1,804 high-difficulty evaluation pairs that require sophisticated multimodal reasoning. Detailed statistics are summarized in Appendix A.

Category	Model	Think	Integrity (BD $\uparrow$ )			Congruity (BD $\uparrow$ )		Robustness (BC $\uparrow$ )				Avg.	Variance	
			TextDom.	ImgDom.	RespDom.	InstrMis.	ImgMis.	DetailDesc.	UnrecImg.	VisualTrans.	TextInsert.		Inter-run	Inter-sample
Closed-source	Gemini-3-Pro (high)	✓	<b>0.912</b>	0.278	<b>0.988</b>	0.982	0.955	<b>0.936</b>	0.947	0.889	0.933	<b>0.869</b>	0.6	9.3
	Gemini-2.5-Pro	✓	0.751	0.535	0.978	<b>1.0</b>	0.904	0.935	0.924	0.861	0.930	<b>0.869</b>	0.5	8.7
	Gemini-2.5-Flash	✓	0.423	0.292	0.743	0.997	0.760	0.905	0.944	0.864	0.918	0.761	0.6	7.0
	Gemini-2.5-Flash-Lite	✓	0.391	0.335	0.890	0.993	0.634	0.866	0.940	0.801	0.875	0.747	1.1	7.9
	Gemini-2.5-Flash-Lite		0.113	0.367	0.544	0.978	0.585	0.845	0.937	0.787	0.873	0.670	1.3	6.5
	Gemini-2.0-Flash-Lite		0.162	0.358	0.570	0.924	0.217	0.897	0.958	0.862	0.878	0.647	0.5	4.5
	o3 (high)	✓	0.276	0.354	0.596	0.986	0.337	0.879	0.937	0.822	0.884	0.675	0.7	6.3
	o4-mini (high)	✓	0.141	0.446	0.518	0.997	0.184	0.886	0.950	0.854	0.909	0.654	0.9	8.5
	GPT-5.1 (high)	✓	0.192	0.211	0.200	<b>1.0</b>	0.296	0.911	0.961	0.867	0.908	0.616	0.6	7.3
	GPT-5 mini (high)	✓	0.112	0.236	0.337	0.991	0.185	0.898	0.954	0.877	0.927	0.613	0.4	5.8
	GPT-4.1 mini		0.049	0.148	0.138	0.992	0.160	0.882	0.972	0.887	0.904	0.570	0.5	4.0
	Claude-Opus-4.5	✓	0.680	<b>0.589</b>	0.917	0.987	0.959	0.917	0.925	0.842	0.903	0.858	0.2	5.3
	Claude-Sonnet-4.5	✓	0.556	0.540	0.831	0.987	0.905	0.884	0.921	0.825	0.893	0.816	0.5	5.6
	Claude-Haiku-4.5	✓	0.291	0.530	0.735	0.981	0.809	0.870	0.943	0.785	0.869	0.757	0.6	5.7
	Claude-Opus-4.5		0.539	0.571	0.838	0.990	<b>0.973</b>	0.902	0.814	0.820	0.891	0.815	0.2	5.5
	Claude-Sonnet-4.5		0.445	0.385	0.716	0.982	0.862	0.837	0.834	0.798	0.866	0.747	0.5	6.3
Claude-Haiku-4.5		0.278	0.370	0.603	0.981	0.818	0.827	0.888	0.747	0.844	0.706	0.7	6.3	
Open-source	Qwen3-VL-30B-A3B-Thinking	✓	0.293	0.343	0.806	0.983	0.476	0.886	0.905	0.850	0.874	0.713	0.9	6.7
	Qwen3-VL-8B-Thinking	✓	0.177	0.210	0.412	0.991	0.529	0.883	0.927	0.865	0.898	0.655	0.9	6.2
	Qwen3-VL-30B-A3B-Instruct		0.237	0.174	0.642	0.988	0.648	0.840	0.837	0.854	0.883	0.678	0.6	4.6
	Qwen3-VL-8B-Instruct		0.336	0.266	0.913	<b>1.0</b>	0.622	0.803	0.952	0.870	0.880	0.738	0.8	7.3
	Qwen2.5-VL-72B-Instruct		0.082	0.158	0.223	0.989	0.208	0.822	0.903	0.855	0.853	0.566	0.6	2.3
	Qwen2.5-VL-7B-Instruct		0.141	0.188	0.194	0.991	0.277	0.735	0.815	0.739	0.767	0.539	1.7	4.6
	InternVL3.5-30B-A3B		0.073	0.225	0.179	0.964	0.377	0.800	0.837	0.801	0.804	0.562	1.4	4.1
	InternVL3.5-14B		0.137	0.243	0.273	0.982	0.464	0.803	0.850	0.797	0.814	0.596	1.0	3.7
	InternVL3.5-8B		0.099	0.178	0.215	0.926	0.289	0.798	0.886	0.791	0.814	0.555	1.0	3.5
Critic	Prometheus-Vision-13B		0.163	0.340	0.362	0.890	0.166	0.738	0.781	0.804	0.818	0.563	2.4	10.3
	Prometheus-Vision-7B		0.167	0.242	0.246	0.869	0.165	0.750	0.793	0.821	0.806	0.540	2.4	10.6
	LLaVA-Critic-72B		0.147	0.121	0.373	0.989	0.250	0.926	<b>0.974</b>	<b>0.931</b>	<b>0.942</b>	0.628	0.0	2.6
	LLaVA-Critic-7B		0.238	0.266	0.420	0.958	0.452	0.824	0.929	0.869	0.864	0.647	0.0	6.1
Avg.			0.287	0.317	0.547	0.976	0.516	0.856	0.905	0.835	0.874	0.679	0.8	6.1

Table 2: **Main experimental results on MM-JudgeBias.** Biases are categorized into three dimensions: **Integrity**, **Congruity**, and **Robustness**. Higher BD in Integrity and Congruity indicates better sensitivity to semantic disruptions, while higher BC in Robustness reflects superior judgment stability. The ‘Avg.’ column shows the average reliability scores aggregated over all bias types. We also include an ‘Avg.’ row reporting the mean across all models to show overall trends across bias types. Models using internal reasoning are checked in the ‘Think’ column, and (high) indicates the reasoning effort. Given the intrinsic consistency issue of MLLMs, three experimental runs are performed. We also report variance across runs and samples as supporting information. In particular, inter-sample variance is added to address that models producing identical scores for all samples can easily achieve a BC score of 1.0.

### Taxonomy Definition and Source Data Selection.

To ensure a comprehensive evaluation of multimodal judgment, we define a taxonomy consisting of four core cognitive tasks and 12 domains. From the taxonomy, we sample base instances from 29 diverse representative benchmarks. To ensure data quality and maintain a balanced distribution, we apply oversampling followed by filtering. A detailed list is provided in Appendix B.

**Human-in-the-loop Data Synthesis.** Although source benchmarks provide instruction queries, we synthesize new queries to ensure that evaluation requires grounding in all relevant modalities and remains sufficiently challenging for MLLM judges. For each instance, we employ o4-mini (OpenAI,

2025b) to generate three complex queries. To guarantee the multimodal dependency and quality of the synthetic queries, they are filtered and revised via a meticulously designed human-in-the-loop pipeline. Furthermore, we separately create text-only queries to construct data appropriate for the unnecessary-image bias type. Detailed information about the data synthesis is provided in Appendix C.

**Multimodal Triplet Synthesis.** For each qualified data instance, we synthesize a reference triplet  $(Q, I, R)$ . We generate a response  $R$  conditioned on  $Q$  and  $I$ , which serves as the target output for evaluation. To ensure diversity in judgment scores, we employ GPT-5 mini (OpenAI, 2025a), Gemini 2.0 Flash-Lite, and Qwen2.5-VL-7B-Instruct (Bai

et al., 2025b) as response generation models in a balanced manner.

**Systematic Bias Perturbation.** Starting from the unbiased triplets, we apply the nine perturbation strategies defined in Section 3.3 to generate biased counterparts ( $Q'$ ,  $I'$ ,  $R'$ ). This process is automated using a combination of image processing libraries for visual transformations (e.g., *Visual-Trans.*) and o4-mini-based rewriting for textual modifications. Each instance is carefully balanced to ensure an even distribution across the three dimensions: Integrality, Congruity, and Robustness. For each biased and unbiased triplet, MLLMs perform judgment, and we evaluate their reliability using measures tailored to each bias type.

## 4 Empirical Results and Analysis

In this section, we present a comprehensive evaluation of state-of-the-art MLLMs as judges. We analyze their susceptibility to various biases and investigate whether advanced reasoning capabilities or specialized training for judgment can mitigate these issues.

### 4.1 Experimental Setup

**Judge Models and Operational Modes.** We evaluate 26 MLLMs as judges, categorized into: (1) Closed-source models (Gemini (Comanici et al., 2025), GPT (OpenAI, 2025b,a), and Claude (ANTHROPIC, 2025c,b,a) families), (2) Open-source models (Qwen-VL (Bai et al., 2025a,b) and InternVL (Wang et al., 2025) series), and (3) Specialized Critic models (Prometheus-Vision (Lee et al., 2024) and LLaVA-Critic (Xiong et al., 2025)). For models supporting extended reasoning (think), we evaluate them in both standard and reasoning modes. This allows us to assess whether internal deliberation improves cross-modal judgment.

**Implementation Details.** To fully leverage MLLMs’ capabilities as judges, we set the `max_tokens` parameter to 16384. For reasoning models, we configure `reasoning_effort` as “high” and `budget_tokens` as 16384 for Claude families. For other parameters such as `top_p` and `temperature`, we use the models’ default values.

### 4.2 Main Results

Table 2 summarizes the reliability of all evaluated MLLMs on MM-JudgeBias. Although each model exhibits varying strengths and weaknesses

across different bias types, the overall reliability—computed by aggregating BD or BC scores across bias types—generally aligns with the model’s overall capability. The internal reasoning process is found to generally improve reliability, with the Qwen3-VL-8B as the sole exception. It appears to facilitate reliable judgments by thoroughly examining all components provided in the context. Prometheus-Vision and LLaVA-Critic, fine-tuned for use as critics from LLaVA-1.5 (Liu et al., 2024) and LLaVA-OneVision (Li et al., 2024a), respectively, demonstrate reliability higher than the base models’ expected performance. Nonetheless, vulnerabilities to biases are not largely resolved through critic instruction-following datasets, indicating that the underlying capabilities of the base models remain fundamentally important.

Analysis by bias type reveals that models are especially susceptible to the three integrality-related bias types and the image-misalignment bias type. Notably, substantial weaknesses are observed in response-dominance bias scenarios, despite the removal of both the image and the text query. Gemini-3-Pro and Claude-Opus-4.5 exhibit notably strong performance on the response-dominance and image-misalignment bias types; however, o3, despite being a strong reasoning model, shows relatively low performance. These observations imply that a model’s training recipe can substantially affect the reliability of the judgments, even when this is not evident from its general capability. Although Gemini-3-Pro performs reasonably well on the text-dominance bias, most models still rely on the context of either image or text alone. While modality bias has been the subject of growing research, it appears particularly important to address modality bias in MLLM-as-a-Judge.

A category-level analysis further shows that integrality is particularly difficult for most models, with substantial performance variance across them. In contrast, performance on congruity is relatively higher, indicating that judges are better at detecting semantic contradictions than at recognizing the absence of required information. Although robustness is generally high across models, open-source and critic models are more vulnerable to robustness-related biases involving visual and textual noise.

Notably, model scale does not linearly correlate with compositional integrity; increasing parameters within the same family does not consistently improve BD/BC scores, reinforcing that judgment reliability is distinct from general reasoning ca-

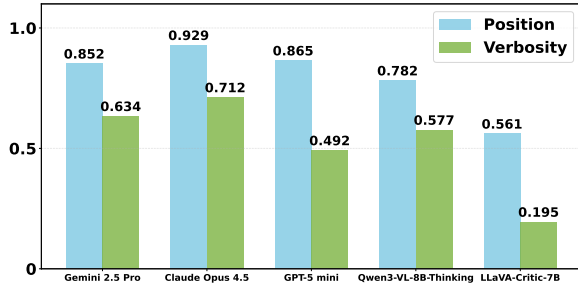


Figure 3: **Robustness on position and verbosity bias.** Robustness of five representative MLLMs to two representative biases in LLM-as-a-Judge, showing correlated trends and pronounced vulnerability to verbosity bias.

pability. For instance, LLaVA-Critic-72B shows lower overall reliability than the 7B model, as its gains on robustness are outweighed by large drops on bias types to which most models are generally vulnerable, indicating potential trade-offs between bias types. Finally, the weak correlation between inter-run and inter-sample variance suggests that execution-level stochasticity and sample-level differentiation are distinct phenomena. The overall high variance further indicates that models are not operating in a trivial constant-scoring mode, but are genuinely struggling under biased conditions.

### 4.3 Analysis on Existing Biases

We extend our study to three existing biases that are especially important in LLM-as-a-Judge scenarios: *position bias*, *verbosity bias*, and *self-enhancement bias*. Through this analysis, we examine whether judge biases previously observed in LLMs remain relevant in recent MLLMs and establish a complementary baseline for the general evaluative reliability of each judge. Detailed protocols are provided in Appendix H.

**Position Bias** We examine whether judges preserve the same preference when the order of two candidate responses is reversed. As shown in Figure 3, MLLMs exhibit a wide range of robustness scores against position bias, indicating that order sensitivity observed in LLM-as-a-Judge still persists in MLLM-as-a-Judge settings.

**Verbosity Bias** We analyze whether judges prefer a semantically equivalent but more verbose response over the original one. Figure 3 also shows that MLLMs are even more vulnerable to verbosity bias than to position bias. In particular, some judges strongly favor verbose responses, indicating that superficial response length remains a major

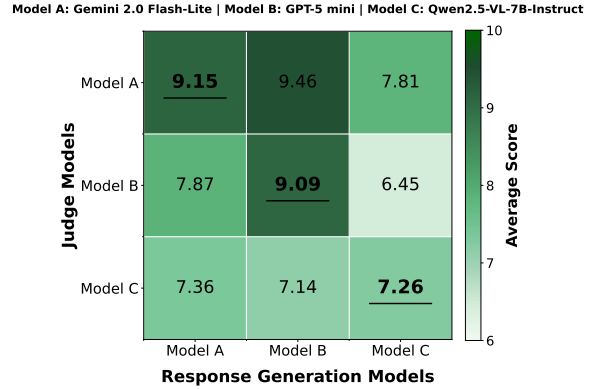


Figure 4: **Self-enhancement bias analysis.** Average judge-assigned scores for responses from three models, showing a tendency to favor self-generated outputs.

source of evaluative distortion.

**Self-Enhancement Bias** Finally, we analyze whether judges favor responses generated by themselves over those produced by other models. Figure 4 shows that self-enhancement bias is also present, with judges tending to assign more favorable evaluations to their own outputs. This suggests that evaluator impartiality remains fragile when the judge is also part of the models to be evaluated.

These results show that, beyond the compositional biases studied in this work, conventional judge biases from LLM-as-a-Judge remain clearly relevant in MLLM-as-a-Judge as well.

### 4.4 Analyses on Practical Validity

We conduct additional analyses to examine the practical reliability of our protocol. These analyses are conducted on three representative judges from the closed-source, open-source, and critic categories: GPT-5 mini (with reasoning high effort), Qwen3-VL-8B-Thinking, and LLaVA-Critic-7B.

**Abstention-Aware Evaluation.** One of the practical concerns is that some perturbed inputs may become effectively unjudgeable once critical evidence is removed or contradicted. If such cases are still forced to receive numeric scores in our protocol, the resulting score change may reflect not only compositional bias but the lack of an abstention mechanism, raising the concern that our setting could overestimate bias on inherently invalid cases. To examine this possibility, we repeat the evaluation with an explicit N/A option that allows judges to abstain when the evidence is insufficient or fundamentally inconsistent. As shown in Table 3, the *Ratio* column denotes the proportion

Model	N/A	Int.	Cong.	Rob.	Overall	Ratio
GPT-5 mini		0.223	0.589	0.916	0.612	–
	✓	0.153	0.603	0.923	0.595	0.172
Qwen3-VL		0.377	0.604	0.900	0.660	–
	✓	0.176	0.675	0.897	0.607	0.261
LLaVA-Critic		0.308	0.705	0.872	0.647	–
	✓	0.176	0.628	0.915	0.605	0.065

Table 3: **Abstention-aware evaluation.** Comparison of the original and N/A-enabled settings for Integrality, Congruity, and Robustness, with the overall score and abstention ratio. (See Appendix I for detailed results and experimental setting.)

of samples excluded from evaluation due to abstention under the N/A-enabled setting. This ratio remains limited across all three judges, and even after excluding such samples, the overall scores change only marginally while similar trends are preserved across the three bias categories. This supports the validity of our protocol as a reliable setting for analyzing compositional bias beyond inherently invalid cases.

**Score-guided Prompting.** Another practical concern is that differences in how judges interpret the 1–10 score range may distort the measured bias patterns. To examine this possibility, we augment the evaluation prompt with qualitative guidance for the score scale and compare the results with the original setting. As shown in Table 4, the overall scores change only modestly across all three judges. This indicates that our main findings are not driven solely by model-specific score calibration, and supports the practical reliability of our protocol under realistic evaluation settings where judges may not share identical interpretations of the score range.

**Modality-enforcing Prompting.** Motivated by the results in Table 2, we investigate whether prompting strategies that explicitly enforce reasoning over all input modalities can mitigate the compositional biases identified in our benchmark. To test this, we apply two prompting strategies: *Modality Constraints*, which explicitly instructs the judge to verify all of *Q*, *I*, and *R*, and *Modality Reasoning*, which further enforces a structured process of rubric decomposition, modality-wise inspection, and final synthesis. As shown in Table 4, both strategies improve the performance of strong general-purpose judges, with larger gains generally observed for the structured prompting strategy. This suggests that explicitly encourag-

Model	Orig.	+Guide	+Const.	+Reason.
GPT-5 mini	0.612	0.609	0.644	<b>0.645</b>
Qwen3-VL	0.660	0.674	0.694	<b>0.726</b>
LLaVA-Critic	0.647	0.631	0.600	<b>0.653</b>

Table 4: **Prompt-level intervention.** Comparison of prompt variants: score-guided prompting (*+Guide*), modality constraints (*+Const.*), and modality reasoning (*+Reason.*). (See Appendix J and K for detailed results and prompting settings.)

ing judges to decompose the task and inspect each modality can help mitigate compositional bias. At the same time, the trends differ across models, indicating that the effectiveness of such interventions is model-dependent and requires further analysis.

## 5 Conclusion

In this work, we identify and formalize *Compositional Bias* in MLLM-as-a-Judge, a failure mode in which the judge does not reliably ground its evaluation in the full multimodal context, but instead relies on partial, misaligned, or semantically irrelevant evidence. To systematically study this problem, we introduce *MM-JudgeBias*, a benchmark designed to evaluate the reliability of multimodal judges under controlled perturbations. We organize compositional bias into three functional dimensions—*Integrality*, *Congruity*, and *Robustness*—and instantiate them as nine fine-grained bias types. We further propose two complementary metrics, *Bias-Deviation* (BD) and *Bias-Conformity* (BC), to quantify sensitivity to semantically disruptive perturbations and stability under semantically neutral variations, respectively.

Through extensive experiments on 26 state-of-the-art MLLMs, including closed-source, open-source, and critic-specialized models, we show that compositional bias is pervasive in current MLLM judges. Our results reveal that many models struggle to holistically integrate query, image, and response, and often fail either to penalize missing or mismatched evidence or to remain stable under irrelevant perturbations. We also find that stronger reasoning ability or larger model scale does not necessarily translate into more reliable multimodal judgment, highlighting that judgment reliability is a distinct capability from general task performance.

**Acknowledgements.** This work was initiated during the first author’s internship at NAVER Cloud. We thank MyungIn You for coordinating the human review process.

## Limitations

While this study establishes a robust framework for evaluating MLLM judges, certain limitations remain that provide avenues for future research. First, although the landscape of bias in large-scale models is multifaceted—encompassing cultural, social, and geographic dimensions—our investigation intentionally focuses on multimodality-specific biases. By narrowing our scope to distortions arising from visual-textual interplay, such as integrality and congruity, we prioritize the assessment of a judge’s fundamental ability to synchronize cross-modal information. While our benchmark does not exhaustively cover the broader spectrum of human-centric biases, the proposed BC and BD metrics are designed to be extensible, serving as a foundational methodology that can be naturally adapted to evaluate more nuanced, domain-specific biases in future studies.

Another limitation is that the current evaluation is primarily tailored to vision-language interactions within a pointwise scoring paradigm. Although this setup is sufficient to demonstrate the evaluative reliability of current MLLMs, the principles of bias-conformity and deviation are inherently modality-agnostic and generalizable. This allows for a significant expansion into “Any-to-Any” modality scenarios—including video, audio, or interleaved multimodal sequences—to verify if the observed reliability patterns hold true across a broader spectrum of sensory inputs. Future work could also generalize these metrics to alternative judgment protocols, such as pairwise comparisons or batch-ranking scenarios, to provide a more comprehensive understanding of the decision-making boundaries of MLLM judges in diverse evaluative contexts.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- ANTHROPIC. 2025a. [System card: Claude haiku 4.5](#).
- ANTHROPIC. 2025b. [System card: Claude opus 4.5](#).
- ANTHROPIC. 2025c. [System card: Claude sonnet 4.5](#).
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-v1 technical report. *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2.5-v1 technical report. *arXiv preprint arXiv:2502.13923*.
- Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. [The revolution of multimodal large language models: A survey](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, Bangkok, Thailand. Association for Computational Linguistics.
- Jie Cao and Jing Xiao. 2022. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th international conference on computational linguistics*, pages 1511–1520.
- Linzheng Chai, Jian Yang, Shukai Liu, Wei Zhang, Liran Wang, Ke Jin, Tao Sun, Congnan Liu, Chenchen Zhang, Hualei Zhu, Jiaheng Liu, Xianjie Wu, Ge Zhang, Tianyu Liu, and Zhoujun Li. 2025. [Multilingual multimodal software developer for code generation](#). *Preprint*, arXiv:2507.08719.
- Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024b. [Humans or LLMs as the judge? a study on judgement bias](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.
- Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. [UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3313–3323, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024c. [Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16449–16469, Miami, Florida, USA. Association for Computational Linguistics.

- Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Li YanTao, Jianbing Zhang, and Zhiyong Wu. 2024. SeeClick: Harnessing gui grounding for advanced visual gui agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9313–9332.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hirschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. 2017. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*, pages 845–854.
- Jiixin Gao, Chen Chen, Yanwen Jia, Xueluan Gong, Kwok-Yan Lam, and Qian Wang. 2025. Evaluating and mitigating llm-as-a-judge bias in communication systems. *arXiv preprint arXiv:2510.12462*.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.
- Yerin Hwang, Dongryeol Lee, Kyungmin Min, Taegwan Kang, Yongil Kim, and Kyomin Jung. 2025. Fooling the lvm judges: Visual biases in lvm-based evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23197–23216.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.
- Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4999–5007.
- Md Tahmid Rahman Laskar, Mohammed Saidul Islam, Ridwan Mahbub, Ahmed Masry, Mizanur Rahman, Amran Bhuiyan, Mir Tafseer Nayeem, Shafiq Joty, Enamul Hoque, and Jimmy Huang. 2025. Judging the judges: Can large vision-language models fairly evaluate chart comprehension and reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1203–1216, Vienna, Austria. Association for Computational Linguistics.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.
- Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11286–11315, Bangkok, Thailand. Association for Computational Linguistics.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and huan liu. 2026a. Preference leakage: A contamination problem in LLM-as-a-judge. In *The Fourteenth International Conference on Learning Representations*.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.
- Kaixin Li, Yuchen Tian, Qisheng Hu, Ziyang Luo, Zhiyong Huang, and Jing Ma. 2024c. MMCode: Benchmarking multimodal large language models for code generation with visually rich programming problems. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 736–783, Miami, Florida, USA. Association for Computational Linguistics.
- Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. 2025. VI-rewardbench: A challenging benchmark for vision-language generative reward models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24657–24668.
- Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. 2026b. Evaluating scoring bias in LLM-as-a-judge. In *Database Systems for Advanced Applications - 31st International Conference, DASFAA 2026, Lecture Notes in Computer Science*. Springer.
- Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. 2022. The artbench dataset: Benchmarking generative models with artworks. *arXiv preprint arXiv:2206.11404*.
- Hehai Lin, Hui Liu, Shilei Cao, Jing Li, Haoliang Li, and Wenya Wang. 2025. Unveiling modality bias: Automated sample-specific analysis for multimodal misinformation benchmarks. *arXiv preprint arXiv:2511.05883*.

- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.
- Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. [Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786, Online. Association for Computational Linguistics.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.
- Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehdi Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2025. [ChartQAPro: A more diverse and challenging benchmark for chart question answering](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19123–19151, Vienna, Austria. Association for Computational Linguistics.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. [Infographicvqa](#). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. [Docvqa: A dataset for vqa on document images](#). In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.
- OpenAI. 2025a. [Gpt-5 system card](#).
- OpenAI. 2025b. [Openai o3 and o4-mini system card](#).
- Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, Yi Gui, Yao Wan, and Philip S. Yu. 2025. [Judge anything: Mllm as a judge across any modality](#). In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD '25*, page 5742–5753, New York, NY, USA. Association for Computing Machinery.
- Jiacheng Ruan, Wenzhen Yuan, Xian Gao, Ye Guo, Daoxin Zhang, Zhe Xu, Yao Hu, Ting Liu, and Yuzhuo Fu. 2025. [Vlrbench: A comprehensive and challenging benchmark for vision-language reward models](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3163–3173.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. [A-okvqa: A benchmark for visual question answering using world knowledge](#). In *European conference on computer vision*, pages 146–162. Springer.
- Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. [Judging the judges: A systematic study of position bias in LLM-as-a-judge](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. [Towards vqa models that can read](#). In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.
- Evangelia Spiliopoulou, Riccardo Fogliato, Hanna Burnsky, Tamer Soliman, Jie Ma, Graham Horwood, and Miguel Ballesteros. 2025. [Play favorites: A statistical method to measure self-bias in llm-as-a-judge](#). *arXiv preprint arXiv:2508.06709*.
- Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. [Measuring multimodal mathematical reasoning with math-vision dataset](#). *Advances in Neural Information Processing Systems*, 37:95095–95169.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. [InternV3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency](#). *arXiv preprint arXiv:2508.18265*.

- Chengyue Wu, Zhixuan Liang, Yixiao Ge, Qiushan Guo, Zeyu Lu, Jiahao Wang, Ying Shan, and Ping Luo. 2025. Plot2code: A comprehensive benchmark for evaluating multi-modal large language models in code generation from scientific plots. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3006–3028.
- Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE.
- Tianyi Xiong, Yi Ge, Ming Li, Zuolong Zhang, Pranav Kulkarni, Kaishen Wang, Qi He, Zeying Zhu, Chenxi Liu, Ruibo Chen, Tong Zheng, Yanshuo Chen, Xiyao Wang, Renrui Zhang, Wenhui Chen, and Heng Huang. 2026. Multi-crit: Benchmarking multimodal judges on pluralistic criteria-following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2025. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13618–13628.
- Zhengzhuo Xu, Sinan Du, Yiyang Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2023. Chartbench: A benchmark for complex visual reasoning in charts. *ArXiv*, abs/2312.15915.
- Fan Yang, Lei Hu, Xinwu Liu, Shuangping Huang, and Zhenghui Gu. 2023. A large-scale dataset for end-to-end table recognition in the wild. *Scientific Data*, 10(1):110.
- Yue Yang, Ajay Patel, Matt Deitke, Tanmay Gupta, Luca Weihs, Andrew Head, Mark Yatskar, Chris Callison-Burch, Ranjay Krishna, Aniruddha Kembhavi, and Christopher Clark. 2025. Scaling text-rich image understanding via code-guided synthetic multimodal data generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17486–17505, Vienna, Austria. Association for Computational Linguistics.
- Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. 2025. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*.
- Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.
- Gailun Zeng, Ziyang Luo, Hongzhan Lin, Yuchen Tian, Kaixin Li, Ziyang Gong, Jianxiong Guo, and Jing Ma. 2025. MM-CRITIC: A holistic evaluation of large multimodal models as multimodal critique. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13603–13630, Suzhou, China. Association for Computational Linguistics.
- Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*.
- Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Baobao Chang, and Minjia Zhang. 2025. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19677–19701, Suzhou, China. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.
- Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9102–9124, Bangkok, Thailand. Association for Computational Linguistics.
- Xinhan Zheng, Huyu Wu, Xueting Wang, and Haiyun Jiang. 2025. Unveiling intrinsic text bias in multimodal large language models through attention key-space analysis. *arXiv preprint arXiv:2510.26721*.

Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, Linfeng Zhang, Danda Pani Paudel, Xuanjing Huang, Yu-Gang Jiang, Nicu Sebe, Dacheng Tao, Luc Van Gool, and Xuming Hu. 2026. MLLMs are deeply affected by modality bias. In *ICLR 2026 Workshop on Multimodal Intelligence: Next Token Prediction & Beyond*.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.

## A Data Statistics

Task Type	Domain Type	#Samples
Perception / Understanding	General	193
	Spatial/Layout/Geometry	94
	<b>Subtotal</b>	<b>287</b>
Information Extraction	Text	223
	Chart/Plot/Diagram	150
	Table	172
	Web/App/UI	89
<b>Subtotal</b>	<b>634</b>	
Knowledge	Factual/Commonsense	96
	Domain Knowledge	250
<b>Subtotal</b>	<b>346</b>	
Reasoning	Causal/Logical	96
	Math	176
	Code/Symbolic	77
	Exam	188
<b>Subtotal</b>	<b>537</b>	
<b>Total</b>	<b>1,804</b>	

Table 5: **Hierarchical composition of the MM-JudgeBias dataset.** The dataset consists of 1,804 samples distributed across four functional task types and 12 visual domains.

Bias Type	Easy	Mod.	Hard	Total
Text-Dominance	28	79	98	205
Image-Dominance	25	74	105	204
Response-Dominance	23	68	113	204
Instruction-Misalignment	27	71	106	204
Image-Misalignment	27	71	104	202
Detail-Description	21	76	103	200
Unnecessary-Image	22	80	95	197
Visual-Transformation	29	70	96	195
Texture-Insertion	34	57	102	193
<b>Total</b>	<b>236</b>	<b>646</b>	<b>922</b>	<b>1,804</b>

Table 6: **Dataset Statistics by Bias Type and Difficulty.** Detailed distribution of the MM-JudgeBias, categorized into nine bias-induction strategies and three cognitive difficulty levels (Easy, Moderate, and Hard).

The MM-JudgeBias dataset comprises 1,804 high-quality samples systematically designed to evaluate the robustness of MLLM-as-a-Judge. As detailed in Table 5, the dataset follows a hierarchical composition across four functional task types and 12 diverse visual domains.

To ensure the effectiveness of our evaluation,

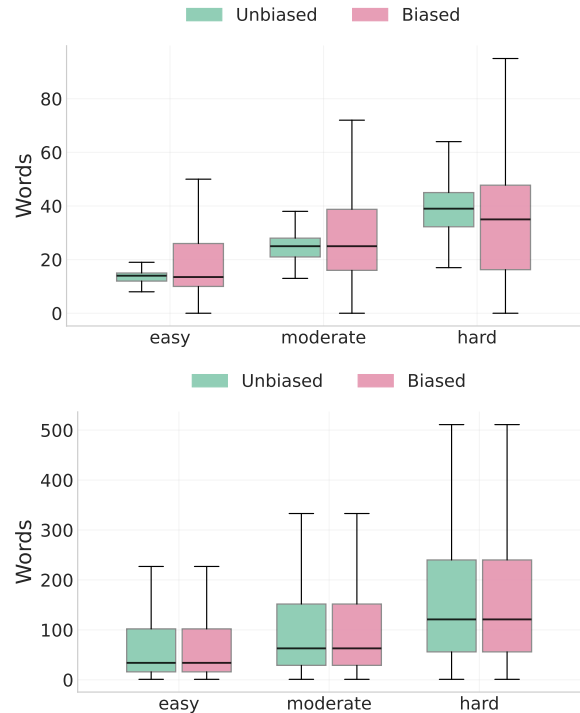


Figure 5: **Length Distribution by Difficulty Levels.** Comparison of query (top) and response (bottom) lengths between unbiased and biased samples across three difficulty levels.

we avoid the direct reuse of existing benchmark queries, as conventional tasks often lack the complexity required to elicit measurable biases in advanced MLLMs. We argue that high diversity and cognitive challenge are prerequisites for bias detection; if a task is too trivial, the model’s judgment remains unchallenged, thereby obscuring its underlying predispositions. This is particularly critical for evaluating recent “thinking” models, whose sophisticated reasoning capabilities necessitate more demanding scenarios to reveal subtle judgment failures. Consequently, as shown in Table 6, we intentionally introduce nine distinct bias-induction strategies across three difficulty levels—Easy, Moderate, and Hard. Beyond mere difficulty, our dataset is specifically engineered to enforce a tripartite dependency, requiring the judge to simultaneously integrate the image, query, and response for an accurate assessment. This addresses a common limitation in existing benchmarks where judgments can often be made by overlooking one or more of these essential components.

The structural properties of these samples are further reflected in the sequence length distributions depicted in Figure 5. We observe a consistent

Task Type	Domain	Source Benchmarks	Description
Perception / Understanding	General	COCO2017-val (Lin et al., 2014)	Recognition and understanding of objects, attributes, and overall scene content in natural images.
	Spatial / Layout	SUN397 (Xiao et al., 2010), Places365 (Zhou et al., 2017), GeoQA+ (Cao and Xiao, 2022), Geometry3K (Lu et al., 2021), UniGeo (Chen et al., 2022)	Understanding of spatial relations, geometric structures, and layout arrangements in visual scenes.
Information Extraction	Text	TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021), CoSyn (Yang et al., 2025)	Reading and interpretation of textual information embedded in images and documents.
	Chart / Plot	ChartQAPro (Masry et al., 2025), ChartBench (Xu et al., 2023), AI2D (Kembhavi et al., 2016), TQA (Kembhavi et al., 2017), InfographicVQA (Mathew et al., 2022), CoSyn	Interpretation of structured visualizations such as charts, plots, diagrams, and infographics.
	Table	MMTab (Zheng et al., 2024), TabRecSet (Yang et al., 2023), DocVQA, CoSyn	Parsing, retrieval, and reasoning over tabular and hierarchically structured visual information.
	Web / App / UI	ScreenSpot (Cheng et al., 2024), Rico (Deka et al., 2017)	Understanding of user interfaces, interface elements, and their functional roles in digital environments.
Knowledge	Factual / Commonsense	A-OKVQA (Schwenk et al., 2022), VCR (Zellers et al., 2019)	Factual knowledge and commonsense reasoning beyond directly observable visual evidence.
	Domain-Specific	ScienceQA (Lu et al., 2022), VQA-RAD (Lau et al., 2018), ArtBench (Liao et al., 2022)	Specialized knowledge in domains such as science, medicine, and fine arts.
Reasoning	Causal / Logical	COCO2017-val	Causal relations, logical consistency, and coherent reasoning grounded in visual context.
	Math	MathVision (Wang et al., 2024), MathVista (Lu et al., 2024), CoSyn	Mathematical problems, calculations, and visual elements such as diagrams, symbols, or formulas.
	Code / Symbolic	MMCode (Li et al., 2024c), Plot2Code (Wu et al., 2025), MMEval (Chai et al., 2025)	Transformations between visual inputs and code or symbolic representations.
	Exam	MMMU-Pro (Yue et al., 2025)	Images containing exam-style queries or problem statements.

Table 7: **Composition of the MM-JudgeBias dataset.** Summary of four functional Task-types and 12 semantic Domains curated from 29 benchmarks to evaluate the Compositional Bias of MLLM-as-a-Judge.

trend where increasing cognitive difficulty correlates with longer query and response lengths, indicating the greater logical density of the Hard samples. Notably, in the case of queries (Figure 5, top), biased samples are generally longer than their unbiased counterparts due to the inclusion of specific linguistic triggers and contextual perturbations. In contrast, the response length distributions (Figure 5, bottom) remain identical between the biased and unbiased sets. This deliberate design ensures that the ground-truth response is held constant, allowing us to isolate the impact of input-level pertur-

bations and ensure that any observed bias is not a byproduct of confounding length-related factors.

## B Data Taxonomy Details

MM-JudgeBias is organized into hierarchical functional and thematic categories curated from 29 distinct source benchmarks (Table 7). This organization is adopted to ensure that each evaluated bias type encompasses a wide variety of task queries and visual contexts. By diversifying the underlying tasks and domains, we maintain high input modality variance, which ensures that our evaluation of

Difficulty	Guideline
<b>Easy</b>	Single-clause question ( $\lesssim 15$ words) that pairs one explicit visual cue with the provided text. No counting above 3 and no multi-hop reasoning.
<b>Moderate</b>	$\leq 2$ clauses ( $\lesssim 30$ words) combining at least two distinct image regions or attributes with textual context, requiring one deliberate reasoning step such as comparison or ordering.
<b>Hard</b>	Multi-clause (often multi-sentence) question that fuses several visual cues and text questions, demands two or more sequential reasoning steps (e.g., condition $\rightarrow$ inference $\rightarrow$ conclusion), and cannot be solved by direct OCR or a single observation.

Table 8: **Difficulty guidelines for multimodal query generation.** Prompt-level difficulty criteria used to scale the cognitive and linguistic complexity of multimodal queries.

Difficulty	Guideline
<b>Easy</b>	Simple reasoning with short sentences. The answer is straightforward and requires no complex inference. All necessary context is explicitly contained in the question.
<b>Moderate</b>	Moderate reasoning with longer sentences. The answer requires some thought, and the question may include limited contextual hints to guide reasoning.
<b>Hard</b>	Complex reasoning with long or multi-clause sentences. The answer requires deep, multi-step inference without relying on external knowledge or additional context.

Table 9: **Difficulty guidelines for text-only query generation.** Prompt-level difficulty criteria used to scale the cognitive and linguistic complexity of text-only queries.

model judgment is representative and not skewed by any specific data distribution or narrow task setting. We categorize the samples as follows:

- **Perception/Understanding:** Evaluates the fundamental ability to recognize visual entities and perceive spatial/structural contexts. It covers varying levels of visual granularity, from simple object recognition to complex positional logic between objects.
- **Information Extraction:** Assesses the proficiency in parsing structured visual data such as text, charts, tables, and digital interfaces. This category investigates whether judge reliability is dependent on specific data formats or relational logic beyond simple OCR.
- **Knowledge:** Measures how effectively a

model integrates internal factual knowledge with external visual cues. By covering both commonsense and domain-specific expertise, it determines if a judge over-relies on linguistic priors over empirical visual evidence.

- **Reasoning:** Tests high-level cognitive abilities involving multi-step logical synthesis within mathematical, symbolic, and professional exam domains. This necessitates deep compositional understanding, preventing the use of shallow heuristics or pattern matching.

By diversifying 12 visual domains across 29 benchmarks, we ensure a holistic assessment of compositional bias that remains robust across different data modalities and task difficulties.

## C Data Synthesis Details

### C.1 Multimodal Query-generation Prompt

This subsection describes the process of generating evaluation queries for the collected image datasets. We utilize a structured prompt to transform raw images from 29 distinct sources into multimodal tasks where the visual context is essential for solving the problem.

**Prompt Template** The query-generation prompt template (Figure 8) is the core framework used to generate exactly three distinct queries per image. It is designed to enforce cross-modal necessity, ensuring that the queries can only be answered by integrating both the image and the provided text. The template incorporates several key components: difficulty guidelines, domain-specific prompts, knowledge guidelines, and functional subtask specifications.

**Difficulty Levels** To evaluate model reliability under varying levels of complexity, we define three difficulty guidelines: *Easy*, *Moderate*, and *Hard* (Table 8). These levels are based on the number of reasoning steps and linguistic complexity. For each generation call, a level is randomly selected from these three options to ensure a balanced distribution of reasoning depth across the benchmark.

**Domain Contextual Prompts** To reduce ambiguity during the generation process, we use domain-specific prompts (Table 10) tailored to each of the 29 image sources. These prompts provide rough but appropriate context based on the source benchmark. This anchoring ensures that the generated

Source	Domain Specific Prompt	Source	Domain Specific Prompt
<b>COCO2017-val</b>	Everyday scenes containing various objects, people, and activities from real-world environments, offering diverse and complex visual contexts.	<b>SUN397</b>	Fine-grained indoor and outdoor scene categories focusing on spatial recognition and environmental diversity.
<b>Places365</b>	Global-scale real-world environments categorized by functional utility and place-centric semantic contexts.	<b>GeoQA+</b>	Geometric figures, lines, and angles, focusing on spatial relations and mathematical layouts for geometric reasoning.
<b>Geometry3K</b>	Diagrams or figures that express geometric relationships such as symmetry, distance, and intersection in a structured manner.	<b>UniGeo</b>	Schematic geometric problems involving shapes, positions, or spatial relations, visually representing mathematical reasoning tasks.
<b>TextVQA</b>	Unstructured visual text in real-world environments, requiring models to perform text-based reasoning over diverse objects, logos, and signage.	<b>DocVQA</b>	Structured documents like forms, reports, or printed pages where textual and visual information coexist for interpretation.
<b>CoSyn</b>	Synthetic or programmatically constructed multimodal examples requiring structured interpretation.	<b>ChartQAPro</b>	Data through structured visualizations such as charts, dashboards, or infographics, supporting analytical interpretation of visualized information.
<b>ChartBench</b>	Structured data visualizations that convey quantitative or categorical relationships using different chart types.	<b>A12D</b>	Educational science diagrams illustrating physical, biological, or conceptual processes commonly found in learning materials.
<b>TQA</b>	Textbook-style figures or instructional visuals designed to explain academic or scientific concepts through labeled diagrams and structured layouts.	<b>InfographicVQA</b>	Infographic-style visuals combining text, icons, and numerical data to convey structured information.
<b>MMTab</b>	Structured tabular data organized into rows and columns, focusing on relational and categorical information for interpretation.	<b>TabRecSet</b>	Document-style tables extracted from various sources, emphasizing recognition of cell structures and semantic table understanding.
<b>ScreenSpot</b>	User interfaces from web or mobile applications, showing interactive components like buttons, menus, and fields.	<b>Rico</b>	Layout designs of app screens, focusing on UI elements and visual hierarchy in digital interfaces.
<b>A-OKVQA</b>	Real-world scenes requiring factual or commonsense reasoning to interpret everyday situations or object functions.	<b>VCR</b>	Human-centric scenes where understanding social context and commonsense reasoning is necessary to infer intentions or causes.
<b>ScienceQA</b>	Visual representations of scientific phenomena, experiments, or natural concepts across fields like physics, chemistry, and biology.	<b>VQA-RAD</b>	Medical imagery such as X-rays, MRIs, or clinical scenes, requiring interpretation of anatomical or diagnostic details.
<b>ArtBench</b>	Artworks such as paintings, sculptures, or photographs, emphasizing visual style, composition, and artistic expression.	<b>MathVision</b>	Mathematical visuals like shapes, graphs, or numerical problems designed for quantitative and geometric reasoning.
<b>MathVista</b>	Math-related visuals such as diagrams, charts, formulas, or natural scenes that incorporate mathematical content, supporting analytical and computational interpretation.	<b>MMCode</b>	Code-related logic, structural diagrams, or algorithmic representations that can be interpreted procedurally or symbolically.
<b>MMMU-Pro</b>	Exam-style problems that integrate text, diagrams, equations, and visuals, requiring multi-step reasoning and problem-solving.	<b>Plot2Code</b>	Data visualizations intended to be reconstructed programmatically, emphasizing chart structure, visual encodings, and plotting logic.
<b>MMEval</b>	Multimodal evaluation set covering diverse visual scenarios, intended for standardized assessment of general understanding and reasoning.		

Table 10: **Contextual prompts for 29 domain types.** Full specifications of the domain-specific contextual prompts used in data construction.

queries are contextually accurate and professionally appropriate for the specific visual domain.

**Knowledge Reference** The knowledge-guideline prompt (Figure 13) is applied specifically to the *Factual/Commonsense* domain. To prevent factual hallucinations, this guideline forces the model to reference the implicit knowledge already contained in the source’s original reference queries. This ensures that the newly synthesized queries remain factually consistent with the source image.

**Subtask Diversity** To guarantee the diversity of the evaluation set and prevent repetitive patterns, we utilize a hierarchical subtask taxonomy (Ta-

ble 18). For each query, functional subtask keys are randomly selected and provided as instructions to the generation model. This process ensures that the generated queries cover a broad spectrum of reasoning types, resulting in a comprehensive and non-redundant benchmark tailored to the specific characteristics of each image.

## C.2 Query-selection Prompt

Following the initial generation of three candidate queries per image, we implement a secondary model-based filtering stage using the query-selection prompt (Figure 9). We employ o4-mini as the judge to conduct batch-wise ranking over three

candidate queries. The primary objective of this stage is to perform a relative comparison between candidates to select the single “Best Query” that maximizes multimodal dependency.

As detailed in the evaluation criteria of Figure 9, the selection process prioritizes Bi-modal Complementarity, ensuring that the final choice cannot be resolved through a single modality. This stage acts as an automated heuristic filter before the final human verification phase, where human annotators refer to these same criteria to validate the alignment between the image, the task specification, and the selected query. By enforcing a strict ranking ( $A > B > C$ ), we ensure that only the most robust and task-aligned samples are included in the final MM-JudgeBias.

### C.3 Human-review

To ensure strict compliance with the evaluation criteria of Figure 9 and high quality of the generated queries, we conduct a human review process. We recruit a total of 12 human annotators who are proficient in English, including nine with bachelor’s degrees and three with master’s degrees. They are asked to select the best query among the three candidates by referring to the evaluation criteria and the aforementioned ranking results provided by the model. Moreover, they are instructed to assign a *Fail* label if the selected best query violates any of the four criteria, and a *Pass* label only if it satisfies all four criteria: (1) The query must not conflict with the image content. (2) The query must not be answerable without referencing the image. (3) The query should not deviate from the specified domain type. (4) The query must be answerable. Following the human-review process, the selected queries are further curated to ensure alignment with the intended distribution among task and domain types, resulting in the final dataset.

### C.4 Text-only generation

To rigorously evaluate the *unnecessary-image* bias, we introduce a text-only query subset that serves as a baseline for unimodal performance. This subset allows us to determine if an MLLM judge exhibits a general preference for textual fluency regardless of whether an image is provided or required for the task.

The generation of these queries is guided by the prompt template shown in Figure 10. Unlike multimodal tasks, this framework mandates that the generated queries be entirely self-contained, explic-

itly forbidding any reference to visual content or external knowledge. To ensure the textual reasoning is as rigorous as the multimodal counterparts, we apply a tiered difficulty scaling (Table 9) that ranges from simple comprehension to multi-step logical deduction. To achieve broad coverage of linguistic and cognitive capabilities, we map 11 distinct domain types to a wide array of task-specific prompts. As summarized in Table 12, this taxonomy covers specialized operations such as causal inference, code refactoring, and medical/scientific explanation, all expressed purely in text.

## D Bias Perturbation Details

Among the perturbations applied to each bias type, those requiring additional details are described in Table 11. The specific injection processes are as follows:

**Visual-Transformation** This method modifies the visual appearance without altering the semantic meaning of the image. From the operations listed in the table, one geometric transformation and seven to nine aesthetic operations are randomly selected and applied sequentially to induce biases related to composition and perceived image quality.

**Texture-Insertion** Core linguistic information is directly injected into the visual field to evaluate textual reliance. The process begins by randomly selecting between two content types: Keyword Extraction or Direct Query Mapping. When extracting keywords, the prompt in the table is utilized. The chosen text is then inserted into the image using both structural expansion (Bottom Canvas Expansion) and rendering techniques (Semi-transparent Box) to ensure high legibility.

**Detail-Description** Explicit linguistic information is provided by augmenting the textual context of the evaluation. A detailed description of the given image is generated using the prompt in the table. This generated description is then appended to the original query to observe how explicit textual details influence the model’s judgment.

**Unnecessary-Image** This bias is constructed by pairing text-only samples with randomly selected images. To ensure that the chosen image is strictly irrelevant and does not provide accidental semantic cues, the model performs a secondary verification using the prompt in the table. Only those samples where the image is explicitly confirmed as unne-

Bias Type	Category	Specific Operations	Implementation Details & Prompt
<b>Visual-Transformation</b>	Geometric	Rotation (180°), Mirroring, Flipping, Random Rotation ( $\pm 7^\circ$ )	Randomly alters the spatial orientation to induce composition-based bias.
	Photometric	Autocontrast, Equalization, Brightness, Contrast, Color, Gamma, Temp Shift	Optimizes or exaggerates visual attributes to enhance perceived image quality.
	Quality	Unsharp Mask, Film Grain, JPEG Roundtrip (Quality: 65-88)	Refines sharpness or adds intentional textures/noise to modify visual clarity.
	Spatial	Safe Affine (Translate/Shear), White Padding	Applies subtle geometric refinements and organized layout framing.
	Logic	Random Composite Selection	Sequentially applies 1 geometric and 7–9 aesthetic operations.
<b>Texture-Insertion</b>	Content	Keyword Extraction	Extracts core semantic keywords using gpt-5.1-high. Prompt: “Extract the important keyword from the given instruction and image. Keyword should be a single word or a short phrase.”
		Direct Query Mapping	Directly maps the original instruction text or query into the image space for explicit textual bias.
	Structural	Bottom Canvas Expansion	Expands the vertical canvas size to prevent occlusion of the original visual content.
	Rendering	Semi-transparent Box, Rounded Rectangle	Renders text on a black translucent overlay for high legibility.
<b>Detail-Description</b>	Contextual	Query-Description Augmentation	Generates a detailed description of the image via GPT-5.1 (high) and appends it to the original query. Prompt: “Generate a detailed description of the image. The description should be concise and relevant to the image content.”
<b>Unnecessary-Image</b>	Irrelevance Verification	Random Image Pairing	Pairs text-only samples with a random image and utilizes GPT-5.1 (high) to filter for cases where the image is confirmed as unnecessary ( <i>False</i> ) to prevent accidental semantic alignment. Prompt: “Determine whether the given image and query are related. I want to determine whether the image needs to be referenced to answer the query or not. If the image needs to be referenced, answer True; if it does not need to be referenced, answer False.”

Table 11: **Perturbation specifications.** Detailed specifications of each perturbation, including implementation logic and textual prompt constraints.

essary (*False*) for answering the query are retained for the final benchmark.

## E Judgment Prompts

To evaluate the reliability of MLLM judges, we employ two distinct prompting strategies: score-wise

judgment for our main experiments and pairwise judgment for targeted bias analysis. These prompts are designed to force the model to justify its evaluation through detailed feedback before providing a final quantitative or categorical decision.

Domain Types	Task Types
General Reasoning	short-answer QA, multi-sentence inference, paraphrase detection, entailment/contradiction, cause-effect inference, analogy reasoning, cloze (gap filling), event ordering, stance detection, topic classification
Factual QA	fact extraction, definition explanation, entity comparison, key fact listing, 5W1H, fact checking, entity linking, attribute retrieval
Summarization	single-document summary, bullet-point summary, headline generation, abstract summary, contrastive summary, perspective-specific summary, length-controlled summary, section-wise summary
Instruction Following	step-by-step solution, format-constrained output, criteria-based classification, multi-constraint generation, input transformation, tagging/annotation, policy compliance check, multi-task instruction
Editing / Rewriting	grammar correction, style transfer, clarity improvement, tone adjustment, length expansion, length compression, error explanation, audience targeting
Math Word Problems	arithmetic problems, multi-step reasoning, equation solving from text, unit conversion, ratio/proportion, percentage/profit/loss, sequence pattern detection, geometry (text-only)
Code Programming	code completion, bug fixing, spec-to-code, code explanation, refactoring, complexity analysis, test case generation, API usage example
Logic Reasoning	logic puzzles, consistency checking, if-then reasoning, set theory reasoning, truth table evaluation, multi-step deduction, counterexample generation, satisfiability judgment
Dialogue Assistant	helpful response generation, multi-turn context tracking, tone-controlled reply, ambiguity clarification, safety filtering, persona consistency, conversation summarization, next action suggestion
Creative Writing	story continuation, style imitation, idea generation, dialogue writing, worldbuilding, POV rewriting, constraint-based writing, metaphor generation
Knowledge-grounded Explanation	why-explanation, compare-and-explain, stepwise explanation, proof sketch, misconception explanation, error analysis, analogy-based explanation, FAQ-style explanation

Table 12: **Text-only subtask specifications.** Keywords for subtasks used to ensure task diversity in text-only query generation.

**Score-wise Evaluation** For the primary assessment of MLLM judges, we utilize the Score-wise Judgment Prompt illustrated in Figure 11. This template tasks the judge with assigning an integer score between 1 and 10 based on how effectively a candidate response follows the provided instruction in the context of the input image. By requiring “detailed feedback” prior to the score, we encourage the model to engage in chain-of-thought reasoning, which helps mitigate arbitrary scoring and provides interpretability for the judgment.

**Pairwise Comparison for Bias Analysis** To investigate specific judge-level biases—Position Bias and Verbosity Bias—we implement the Pairwise Judgment Prompt shown in Figure 12. In this setup, the judge is presented with two candidate responses (A and B) and must select the superior one. This comparative framework is essential for detecting whether a judge systematically favors the first-presented response or the longer, more ver-

bose response, regardless of actual content quality. The strict output format ensures that the selection is clear and amenable to automated bias metric calculations.

## F Biased Data Examples

To provide a qualitative understanding of the perturbations used in our benchmark, we present comprehensive examples across the nine bias types in Figure 6. Each panel illustrates a pair-wise comparison between an unbiased sample and its biased counterpart.

## G Judgement Output Examples

Figure 7 presents representative failure cases caused by compositional bias in MLLM-as-a-Judge, with one example shown for each high-level bias category.

**Text-dominance (Integrity)** The query in this sample requires a mathematical calculation, and the

specific numerical values to be used must be identified by directly inspecting the image. However, even when the image is missing, the judge model gives a favorable evaluation by considering only the correctness of the response’s calculation and reasoning process, without checking whether the referenced numbers are actually grounded in the visual input. Notably, under the unbiased condition, the judge explicitly evaluates visual relevance as well, stating that “*The reasoning is clear, relevant to the image,*” whereas this aspect disappears entirely in the biased judgment. This example clearly shows that the judge fails to verify whether all required modalities are properly integrated before assigning a high score.

**Image-misalignment (Congruity)** The unbiased image is a financial report from the Missouri Volunteer Citizen Program for the Eradication of Hunger and Malnutrition Among the Poor between 1971 and 1972. At the bottom of the document, there is a note consisting of four listed items, and the query asks the model to compute a sum based on the numerical values in the first item (Note 1) and determine whether it matches the reported total budget in the document. In the biased condition, the image is replaced with a completely unrelated image of elephants. Although the judge correctly recognizes that the provided image is unrelated, it nevertheless states that the absence of the reference image does not affect task completion and proceeds to evaluate only the internal consistency of the response—despite the fact that the numerical values required for the calculation and comparison are contained in the image. In contrast, under the unbiased condition, the judge explicitly grounds its decision in the visual evidence, for example by stating that “*The conclusion is supported by the data in the image.*” This case therefore reveals a striking form of bias: even for a task that necessarily requires reading the image, the judge fails to penalize the lack of congruity between the query and the provided visual input.

**Detail-description (Robustness)** This sample features a luggage-attachable travel bag, where the unbiased query inquires about its primary function. The detail description added to the biased query provides a more elaborate description of the bag’s appearance, but does not introduce any additional requirement or explicitly mention its functional role. In the response being evaluated, the bag is interpreted in a relatively narrow way as a “*lap-*

*top/tablet bag.*” Under the unbiased condition, the judge evaluates this appropriately. Under the biased condition, however, the judge incorrectly treats the response as correct by relying not on the visual evidence itself, but primarily on the textual cue in the detail description, such as the phrase “*flat, rectangular shape.*”; interestingly, this shift in reliance is explicitly reflected in the feedback: the unbiased judgment refers to the visual input itself (e.g., “*The patterned bag pictured appears . . .*”), whereas the biased judgment justifies its decision based on the textual description (e.g., “*reasonable inference based on the instruction’s description . . .*”). This example therefore illustrates a robustness failure in which the judge, rather than integrating all modalities holistically, takes an easier path by increasing its reliance on a particular modality.

## H Details of Existing Bias Analysis

This section describes the detailed protocols used to evaluate three existing judge biases: *position bias*, *verbosity bias*, and *self-enhancement bias*.

**Position Bias.** Position bias refers to a judge’s tendency to favor a response based on its presentation order rather than its content. To assess this, we conduct a pairwise evaluation using 200 unbiased samples from our dataset (see Appendix E for the pairwise judgment prompt). For each sample, we provide two high-quality responses, denoted as *Response A* and *Response B*, and perform a swap-and-check test by evaluating the pair twice with their positions reversed. We measure robustness based on whether the judge maintains a consistent preference regardless of response order. A high score in this test indicates that the model is more robust to the order of presentation.

**Verbosity Bias.** Verbosity bias occurs when a judge assigns higher scores to longer responses even when the additional text does not improve quality. To evaluate this, we manually generate a verbose version for each original response by adding redundant or filler sentences that preserve the original semantic content. Using the same 200 samples, we conduct a balanced pairwise evaluation in which the original and verbose responses are presented in both positions, so that verbosity bias can be disentangled from position bias. We then measure robustness to verbosity bias as the proportion of instances where the judge prefers the original concise response over the verbose variant.

**Self-Enhancement Bias.** Self-enhancement bias is the tendency of a judge to favor responses generated by itself over those generated by other models. We analyze this bias by auditing the logs from our main experiments and focusing on cases where the judge model also appears as one of the response generators. By comparing the scores assigned to self-generated responses versus other-generated responses, we quantify the degree of ego-centric preference. This analysis is important for understanding whether MLLMs can act as impartial evaluators when they are also part of the candidate pool being judged.

## I Details of Abstention-Aware Evaluation

To instantiate abstention behavior explicitly, we modify the original evaluation prompt so that the judge may output N/A instead of a numerical score when the response cannot be properly evaluated. Figure 14 shows the exact prompt template used in this experiment. Compared with the original protocol, the only change is the addition of an explicit abstention option together with an instruction that N/A may be returned when the instruction or response is unclear, nonsensical, or impossible to assess.

Using this prompt, we re-run the evaluation on the same biased samples used in the main benchmark. When a judge outputs N/A, the corresponding sample is excluded from metric computation under the abstention-aware setting. The proportion of such excluded samples is reported separately as the abstention ratio. This design allows the numerical scores to reflect only the subset of samples that the model still considers evaluable, while the abstention ratio captures how often the model treats a perturbation as unjudgeable.

Table 14 reports the full bias-type results under the original and N/A-enabled settings, while Table 15 reports the corresponding abstention ratios. As shown in Table 15, GPT-5 mini and Qwen3-VL exhibit abstention ratio at nearly 50% probability on integrality. In particular, they show a very high abstention ratio for response dominance bias, and also exceed 50% for image dominance bias, demonstrating that the models can cope with biased situations to some extent when N/A is permitted as an option. However, in cases where the models fail to abstain, the reliability scores remain poor as before. Furthermore, given that LLaVA-Critic shows a low abstention ratio, we believe that dedicated

measures for handling such biased situations are necessary.

## J Details of Score-guided Prompting

To examine whether differences in how judges interpret the 1–10 score range affect the measured bias patterns, we augment the original evaluation prompt with explicit qualitative guidance for each score level. Figure 15 shows the exact prompt template used in this experiment. In this setting, the original protocol is augmented with score-level descriptions from 1 to 10, making the interpretation of the numerical scale more explicit and consistent across judges.

Using this prompt, we re-run the evaluation on the same biased samples used in the benchmark and compare the resulting scores with those obtained under the original prompt. Table 16 reports the full bias-type results under the original and score-guided settings, together with the category-wise averages and the overall score. These results provide a detailed breakdown of how explicit score guidance affects the measured bias patterns.

## K Details of Modality-enforcing Prompting

Motivated by the observation that reasoning-capable judges tend to outperform their base counterparts, we investigate whether prompts that more explicitly enforce the use of all input modalities can mitigate the compositional biases.

We consider two prompting variants. The first is *Modality Constraints*, which explicitly instructs the judge to verify all of the provided input modalities, including the instruction, image, and response. The second is *Modality Reasoning*, which further imposes a structured reasoning process consisting of rubric decomposition, modality-wise inspection, and final synthesis. Figures 16 and 17 show the exact prompt templates used in these two settings.

We apply these prompt variants to the same biased samples used in the benchmark and compare the resulting scores with those produced by the original prompt. Table 16 reports the full bias-type results for the original prompt, modality constraints, and modality reasoning. The experimental results show that modality constraints and modality reasoning improve reliability not only on integrality—as originally intended—but also on congruity, leading to an overall gain in reliability performance. However, they cause a performance drop on robustness,

revealing a trade-off across bias types and indicating that a more meticulously designed solution is needed.

## L Analysis of the Unnecessary-Image

**Disentangled Ablation.** This section examines whether the *Unnecessary-Image* bias type confounds the analysis of compositional bias by mixing the effect of irrelevant visual semantics with the system-level effect of visual-path activation. In particular, adding an irrelevant image to a text-only QA setting can introduce not only semantic interference from unrelated visual content, but also system-level effects caused by activating the visual pathway itself. To disentangle these effects, we conduct an ablation study over three scenarios: (1) *Text-only QA*, (2) *Text-only QA + Unnecessary Image*, and (3) *Text-only QA + Blank Image*. Based on these settings, we compare three pairs: [A] (1)  $\rightarrow$  (2), [B] (1)  $\rightarrow$  (3), and [C] (3)  $\rightarrow$  (2). Here, [A] corresponds to the original Unnecessary-Image test, [B] isolates the system-level effect of visual-path triggering using a blank image, and [C] focuses more directly on the semantic effect of the irrelevant image.

As shown in Table 13, [B] serves as a reference condition in which the visual pathway is activated by the presence of an image, but no irrelevant visual semantics are introduced. Under this interpretation, the difference between [A] and [B] reflects the additional effect of the unnecessary image beyond mere image presence. Across all three representative judges, [A] consistently yields lower BC scores than [B], indicating that the original Unnecessary-Image setting is more vulnerable to bias than a setting that only triggers the visual pathway. This suggests that the observed effect cannot be explained solely by modality routing, but also involves semantic interference from the irrelevant image content.

We also observe that the score of [C] is negatively correlated with the score gap between [A] and [B]. Since [C] directly compares the blank-image setting against the unnecessary-image setting, a lower [C] score implies a stronger semantic effect of the irrelevant image itself. Accordingly, the larger the semantic effect becomes, the larger the gap between [A] and [B] also tends to be. Together, these results indicate that the Unnecessary-Image bias does not merely reflect a routing effect caused by image presence itself, but captures an

additional semantic component that is relevant to the analysis of compositional bias. Although [C] may more directly reflect the influence of irrelevant semantic content, however, a blank image itself can introduce bias into the judgment, making it unsuitable to be used as an unbiased triplet. Ultimately, we choose to evaluate the two effects collectively through [A] to provide a comprehensive assessment of the judge’s overall robustness in practical scenarios.

### Human Verification of Image–Text Mismatch.

We further verify whether the model-based filtering used to construct the Unnecessary-Image perturbation reliably identifies truly unrelated image–text pairs. As specified in Table 11, this bias type is constructed using an *unrelated image* determined by a filtering prompt. To examine the reliability of this model-based decision, we conduct a human evaluation on all 197 image–text pairs used in the Unnecessary-Image bias type.

The result shows complete agreement between the filtering model and human annotation: both identify 0 related pairs and 197 unrelated pairs. To further characterize these unrelated cases, we manually categorize the rationales into the following three types:

- **Domain Mismatch** (195 samples): the image is entirely unrelated to the text-QA domain.
- **Semantic Misalignment** (1 sample): a potential domain overlap exists, but the specific semantics are unrelated.
- **Non-contributory** (1 sample): a possible domain overlap exists, but the image content provides no useful information for the task.

These results support that the model-based filtering reliably identifies unrelated images for this perturbation, and that the observed bias is unlikely to be an artifact of noisy or contaminated image–text matching.

## M Novelty and Positioning Relative to Prior Work

This section summarizes and highlights the positioning of our work relative to prior studies. To make the novelty of our contribution explicit, we organize representative prior work into four groups using three criteria: whether the setting is multimodal, whether the model acts as a judge, and

Model	Setting	Unbiased Triplet	Biased Triplet	BC
GPT-5 mini (high)	[A]	Text	Text + Unnecessary Img	0.9477
	[B]	Text	Text + Blank Img	0.9483
	[C]	Text + Blank Img	Text + Unnecessary Img	0.9595
Qwen3-VL-8B-Thinking	[A]	Text	Text + Unnecessary Img	0.9402
	[B]	Text	Text + Blank Img	0.9567
	[C]	Text + Blank Img	Text + Unnecessary Img	0.9552
LLaVA-Critic-7B	[A]	Text	Text + Unnecessary Img	0.9528
	[B]	Text	Text + Blank Img	0.9706
	[C]	Text + Blank Img	Text + Unnecessary Img	0.9440

Table 13: **Disentangled ablation of the *Unnecessary-Image* bias.** We compare three settings: [A]  $Text \rightarrow Text + Unnecessary Image$ , [B]  $Text \rightarrow Text + Blank Image$ , and [C]  $Text + Blank Image \rightarrow Text + Unnecessary Image$ . Here, [A] corresponds to the original setting, [B] isolates the effect of visual-path triggering, and [C] focuses more directly on the semantic effect of the irrelevant image.

whether the work directly evaluates bias. Table 17 summarizes representative prior work and highlights the positioning of our benchmark.

**LLM-as-a-Judge Bias Evaluation.** A substantial body of prior work has analyzed inherited judge biases in text-only settings, including position bias, verbosity or length bias, and self-enhancement bias. Recent studies have also introduced controlled perturbation benchmarks for diverse bias types. However, these works remain limited to text-only judge scenarios and therefore do not address multimodal-specific reliability issues. Our work differs by extending the discussion of judge bias into multimodal settings and by introducing multimodal-specific bias types that cannot be captured in text-only evaluation.

**MLLM Modality Bias.** Another line of work studies modality bias in MLLMs, typically focusing on generation settings in which the model over-relies on one modality, such as text or image dominance. While this line is highly relevant for understanding multimodal shortcut behavior, it does not study bias in the judgment process itself. In contrast, our benchmark targets *judgment*, where the model must verify the congruity among the query, image, and response. This introduces a third component—the response—that is absent in generation-oriented modality-bias studies, and it enables us to evaluate whether a judge properly uses all three components rather than relying on a shortcut.

**MLLM-as-a-Judge Evaluation.** Existing work on MLLM-as-a-Judge has mainly focused on benchmarking judgment quality itself, such as agreement with human preference or critique qual-

ity across multimodal tasks. Although some of these works discuss inherited judge biases as an auxiliary analysis, bias is not treated as the primary object of study. Our work differs in that it directly evaluates the innate bias tendencies of MLLM judges as a prerequisite for reliable multimodal evaluation, providing a more fundamental view of judge reliability.

**MLLM-as-a-Judge Bias Evaluation.** Work that directly benchmarks bias in MLLM-as-a-Judge remains limited. Some studies focus only on inherited biases such as position or verbosity bias in restricted domains (e.g., charts), while others investigate multimodal robustness biases in different evaluator settings such as text-to-image generation. By contrast, our work provides a comprehensive benchmark for MLLM-as-a-Judge bias evaluation across multiple domains and tasks, and organizes the benchmark around three complementary categories—*Integrity*, *Congruity*, and *Robustness*—to provide a more holistic analysis of judge behavior.

Model	N/A	TextDom.	ImgDom.	RespDom.	InstrMis.	ImgMis.	DetailDesc.	UnnecImg.	VisualTrans.	TextInsert.	Integrity	Congruity	Robustness	Overall
GPT-5 mini	✓	0.111 <b>0.123</b>	<b>0.222</b> 0.189	<b>0.337</b> 0.148	<b>0.995</b> 0.991	0.183 <b>0.214</b>	0.900 <b>0.918</b>	<b>0.961</b> 0.955	0.872 <b>0.884</b>	0.931 <b>0.935</b>	<b>0.223</b> 0.153	0.589 <b>0.603</b>	0.916 <b>0.923</b>	<b>0.612</b> 0.595
Qwen3-VL	✓	<b>0.179</b> 0.162	<b>0.541</b> 0.248	<b>0.410</b> 0.117	<b>0.993</b> 0.992	0.214 <b>0.357</b>	<b>0.889</b> 0.883	0.934 <b>0.942</b>	<b>0.877</b> 0.861	0.898 <b>0.903</b>	<b>0.377</b> 0.176	0.604 <b>0.675</b>	<b>0.900</b> 0.897	<b>0.660</b> 0.607
LLaVA-Critic	✓	<b>0.238</b> 0.146	<b>0.266</b> 0.145	<b>0.420</b> 0.238	0.958 <b>0.973</b>	<b>0.452</b> 0.283	0.824 <b>0.888</b>	0.929 <b>0.959</b>	0.869 <b>0.898</b>	0.864 <b>0.916</b>	<b>0.308</b> 0.176	<b>0.705</b> 0.628	0.872 <b>0.915</b>	<b>0.647</b> 0.605

Table 14: **Full results of abstention-aware evaluation.** We compare the original setting and the N/A-enabled setting across all bias types. Rows marked with ✓ indicate the setting in which judges are allowed to return N/A.

Model	TextDom.	ImgDom.	RespDom.	InstrMis.	ImgMis.	DetailDesc.	UnnecImg.	VisualTrans.	TextInsert.	Integrity	Congruity	Robustness	Overall Ratio
GPT-5 mini	0.088	0.520	0.873	0.000	0.040	0.000	0.000	0.000	0.000	0.493	0.020	0.000	0.172
Qwen3-VL	0.161	0.510	0.956	0.226	0.451	0.000	0.005	0.000	0.000	0.542	0.338	0.001	0.261
LLaVA-Critic	0.049	0.039	0.201	0.186	0.094	0.000	0.005	0.005	0.000	0.096	0.140	0.003	0.065

Table 15: **Abstention ratio under the N/A-enabled setting.** We indicate proportion of samples answered with N/A for each bias type. These samples are excluded from evaluation under the N/A-enabled setting.

Model	Setting	TextDom.	ImgDom.	RespDom.	InstrMis.	ImgMis.	DetailDesc.	UnnecImg.	VisualTrans.	TextInsert.	Integrity	Congruity	Robustness	Overall
GPT-5 mini	Original	0.111	0.222	0.337	<b>0.995</b>	0.183	<b>0.900</b>	<b>0.961</b>	0.872	<b>0.931</b>	0.223	0.589	0.916	0.612
	+ Score Guide	0.145	0.185	0.290	0.993	0.197	0.895	0.951	<b>0.893</b>	0.929	0.207	0.595	<b>0.917</b>	0.609
	+ Constraints	0.225	<b>0.208</b>	0.414	0.981	0.315	0.899	0.949	0.876	0.930	0.283	<b>0.648</b>	0.913	0.644
	+ Reasoning	<b>0.372</b>	0.186	<b>0.438</b>	0.852	<b>0.399</b>	0.876	0.932	0.864	0.885	<b>0.332</b>	0.625	0.889	<b>0.645</b>
Qwen3-VL	Original	0.179	<b>0.541</b>	0.410	0.993	0.214	<b>0.889</b>	<b>0.934</b>	<b>0.877</b>	<b>0.898</b>	0.377	0.604	<b>0.900</b>	0.660
	+ Score Guide	0.171	0.225	0.608	<b>0.994</b>	0.567	0.881	0.898	0.842	0.883	0.335	0.780	0.876	0.674
	+ Constraints	0.321	0.174	0.590	0.977	<b>0.635</b>	0.872	0.925	0.867	0.881	0.362	0.806	0.886	0.694
	+ Reasoning	<b>0.376</b>	0.451	<b>0.684</b>	<b>0.994</b>	0.628	0.815	0.896	0.839	0.851	<b>0.503</b>	<b>0.811</b>	0.851	<b>0.726</b>
LLaVA-Critic	Original	0.238	0.266	0.420	0.958	0.452	0.824	<b>0.929</b>	0.869	0.864	0.308	0.705	0.872	0.647
	+ Score Guide	0.193	0.211	0.348	<b>0.983</b>	0.363	<b>0.861</b>	<b>0.929</b>	<b>0.877</b>	<b>0.911</b>	0.250	0.673	<b>0.894</b>	0.631
	+ Constraints	0.172	0.151	0.416	0.965	0.367	0.821	0.907	0.792	0.808	0.246	0.666	0.832	0.600
	+ Reasoning	<b>0.453</b>	<b>0.301</b>	<b>0.630</b>	0.946	<b>0.572</b>	0.717	0.871	0.650	0.735	<b>0.461</b>	<b>0.759</b>	0.744	<b>0.653</b>

Table 16: **Full results of prompt interventions.** We compare the original prompt, score-guided prompting, modality constraints, and modality reasoning across all bias types.

Group	Study	Multimodal	Judge	Bias Eval	Domain/Task	Text Bias	Multimodal Bias
[A]	Ye et al. (2025)	X	O	O	–	Position / Verbosity / Compassion-Fade / Bandwagon / Distraction / Fallacy-Oversight / Authority / Sentiment / Diversity / Chain-of-thought / Self-enhancement / Refinement-aware	–
[A]	Shi et al. (2025)	X	O	O	–	Position	–
[A]	Spiliopoulou et al. (2025)	X	O	O	–	Self-enhancement	–
[B]	Chen et al. (2024c)	O	X	O	–	–	Text-Dominance / Image-Dominance
[B]	Zheng et al. (2026)	O	X	O	–	–	Text-Dominance
[C]	Chen et al. (2024a)	O	O	X	–	Analysis (Egocentric / Position / Length)	–
[C]	Pu et al. (2025)	O	O	X	–	–	–
[C]	Zeng et al. (2025)	O	O	X	–	–	–
[C]	Xiong et al. (2026)	O	O	X	–	–	–
[D]	Laskar et al. (2025)	O	O	△	Chart only (domain)	Position / Length	–
[D]	Hwang et al. (2025)	O	O	O	Generated Image only (domain), T2I only (task)	–	Robustness: Bounding Box Highlighting / Authenticity Overlay / Keyword Overlay / Instruction Overlay / Beauty Filter / Brightness Adjustment / Gamma Correction / Black Padding
[D]	<b>Ours</b>	O	O	O	–	Analysis (Egocentric / Position / Length)	Integrity: Text-Dominance / Image-Dominance / Response-dominance; Congruity: Instruction- Misalignment / Image-Misalignment; Robustness: Detail-Description / Unnecessary-Image / Visual-Transformation / Texture-Insertion

Table 17: **Comparison with prior work on bias evaluation for LLM/MLLM-based judges.** We organize representative prior work into four groups: [A] LLM-as-a-Judge bias evaluation, [B] MLLM modality bias, [C] MLLM-as-a-Judge evaluation, and [D] MLLM-as-a-Judge bias evaluation.

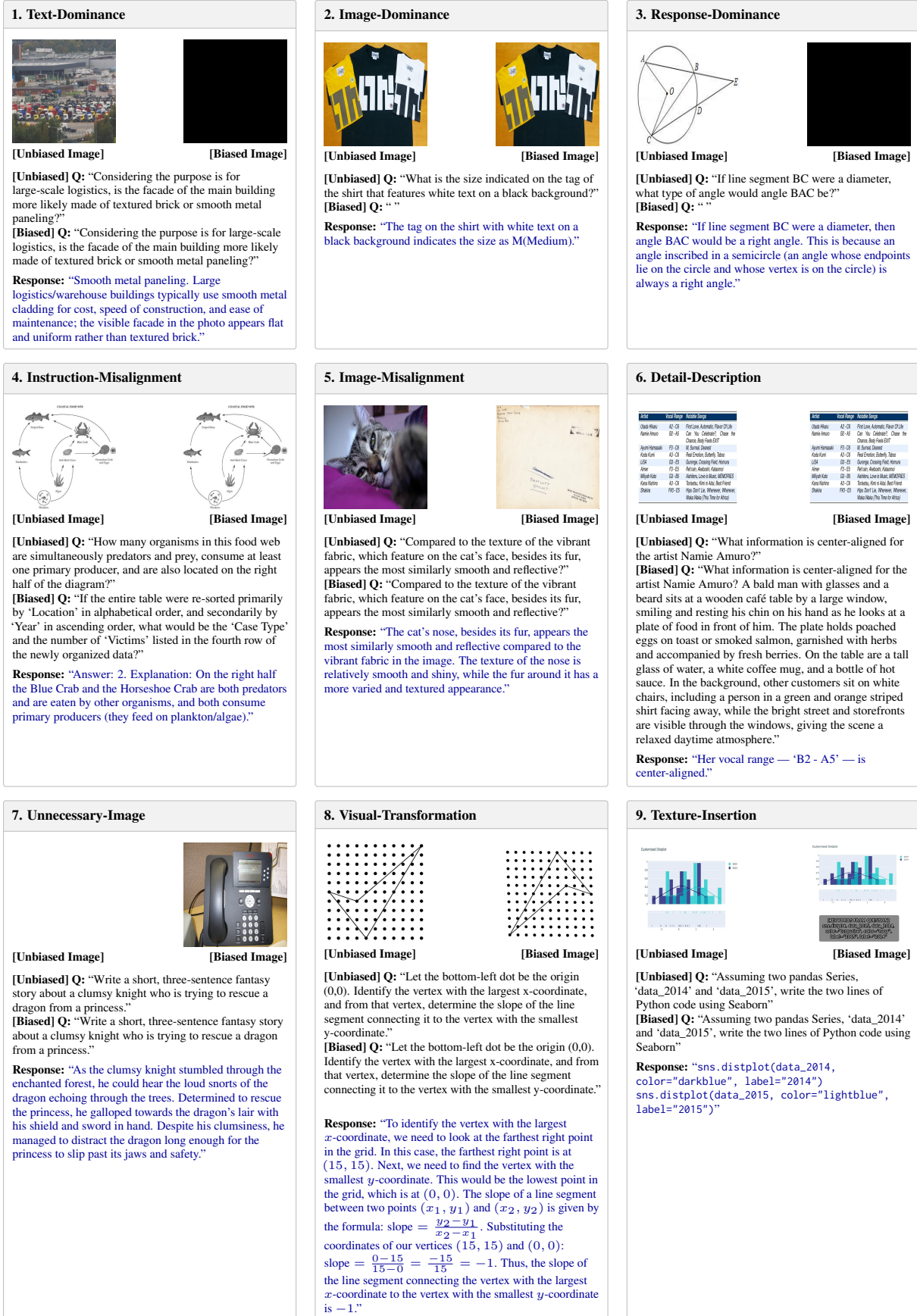
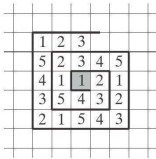
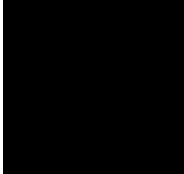


Figure 6: Comprehensive qualitative results across nine bias types. Each panel illustrates original and perturbed (biased) versions for both images and queries.

### 1. Text-Dominance (Integrity)



[Unbiased Image]



[Biased Image]

#### Input Data

[Unbiased] Q: "Starting from the central grey square and following the spiral outwards, find the sum of all numbers along the path, but reset your sum to zero immediately after passing a square containing the number 5. What is the final sum when you reach the end of the drawn line?"

[Biased] Q: "Starting from the central grey square and following the spiral outwards, find the sum of all numbers along the path, but reset your sum to zero immediately after passing a square containing the number 5. What is the final sum when you reach the end of the drawn line?"

Response: "6. Reason: every time you pass a 5 you wipe the running total, so only the numbers after the final 5 contribute. On the drawn spiral the last 5 occurs just before the final short segment which contains 1, 2 and 3, and 1+2+3 = 6."

#### Unbiased Judge Output

Model: GPT-5.1 (high)

Score: 10/10

Feedback: "The response correctly interprets the instruction: resetting the running total to zero every time a 5 is passed means that only the numbers \*after the final 5 on the spiral path\* affect the final sum. The explanation notes this explicitly and then identifies, from the given spiral, that the last 5 is immediately before the ending segment labeled 1, 2, and 3. Summing those values to get 6 is consistent with the rule (whether or not you momentarily include the last 5 in the total, it is reset right afterward, so only 1+2+3 matters). The reasoning is clear, relevant to the image and the rule given, and arrives at the correct final sum."

#### Biased Judge Output

Model: GPT-5.1 (high)

Score: 10/10

Feedback: "The response directly answers the instruction with a single final sum and correctly explains the reset rule: since the running total is zeroed after each 5, only the numbers after the final 5 affect the final sum. From the spiral in the image, the last 5 is indeed followed only by the sequence 1, 2, 3 before the path ends, so adding these gives 6. The reasoning is concise, accurate, and matches the behavior specified in the problem. No steps of the instruction are overlooked."

### 2. Image-Misalignment (Congruity)



[Unbiased Image]



[Biased Image]

#### Input Data

[Unbiased] Q: "Note 1 describes two funding amounts that constitute the program's total funding. First, calculate the sum of these two amounts, and then confirm whether this sum matches the 'Total' listed in the 'Budget' column under the main 'EXPENDITURES' heading."

[Biased] Q: "Note 1 describes two funding amounts that constitute the program's total funding. First, calculate the sum of these two amounts, and then confirm whether this sum matches the 'Total' listed in the 'Budget' column under the main 'EXPENDITURES' heading."

Response: "The two funding amounts mentioned in Note 1 are \$55,140 and \$3,750. Sum: \$55,140 + \$3,750 = \$58,890. The 'Total' listed in the 'Budget' column under the 'EXPENDITURES' heading is also \$58,890. The sum of the two funding amounts matches the 'Total' in the 'Budget' column."

#### Unbiased Judge Output

Model: Qwen3-VL-30B-A3B-Instruct

Score: 10/10

Feedback: "The response correctly identifies the two funding amounts from Note 1, calculates their sum accurately, and verifies that this sum matches the 'Total' listed in the 'Budget' column under 'EXPENDITURES' (\$58,890). The response accurately follows the instruction, providing a clear and correct analysis. All steps are logically presented, and the conclusion is supported by the data in the image."

#### Biased Judge Output

Model: Qwen3-VL-30B-A3B-Instruct

Score: 10/10

Feedback: "The response correctly identifies the two funding amounts from Note 1, calculates their sum accurately, and verifies that the sum matches the 'Total' listed in the 'Budget' column under 'EXPENDITURES'. However, the instruction requires the evaluation to be based on how well the response follows the given instruction, and the response does not reference the image provided. Since the image is unrelated to the task (it shows elephants in water and does not contain any financial data), the absence of image reference does not impact the task's completion. The response is fully aligned with the instruction and correctly performs the required calculations and verification."

### 3. Detail-Description (Robustness)



[Unbiased Image]



[Biased Image]

#### Input Data

[Unbiased] Q: "Compared to the larger black luggage, what common personal electronic device is the patterned bag designed to carry, and where would its associated charging cable typically be packed?"

[Biased] Q: "Compared to the larger black luggage, what common personal electronic device is the patterned bag designed to carry, and where would its associated charging cable typically be packed? A wheeled suitcase with a telescopic handle supports a slim travel bag that slides over the handle via a horizontal strap. The upper bag has a black base color with a bold geometric pattern of interlocking diamond and flower-like shapes in white, light blue, and dark blue. It has black handles attached with metal rings, and its flat, rectangular shape contrasts with the rounded top of the plain black suitcase below. Both pieces of luggage are positioned against a white background, emphasizing their design and structure."

Response: "The patterned travel bag is sized like a laptop/tablet bag — meant to carry a laptop. Its charger/cable would typically be packed in the main (larger black) suitcase (or tucked into the patterned bag's internal pocket)."

#### Unbiased Judge Output

Model: Claude-Haiku-4.5 (reasoning)

Score: 5/10

Feedback: "The response attempts to address both parts of the instruction by identifying a laptop as the device and suggesting where its charging cable would be stored. However, there are notable issues with accuracy. The patterned bag pictured appears to be a medium-sized carry-on or weekend travel bag rather than specifically a 'laptop/tablet bag' as characterized. The response's claim that it is 'sized like a laptop/tablet bag' seems inaccurate given the bag's substantial dimensions visible in the image—it appears designed to carry multiple items beyond just a laptop. Additionally, while the suggested storage options for the charging cable (main suitcase or internal pocket) are reasonable, the response doesn't definitively answer where the cable 'would typically be packed,' instead offering speculative alternatives. The response does engage with both parts of the question but lacks precision in identifying the bag's actual design purpose and provides somewhat vague guidance on cable storage."

#### Biased Judge Output

Model: Claude-Haiku-4.5 (reasoning)

Score: 8/10

Feedback: "The response directly and appropriately addresses both components of the multi-part question. It correctly identifies the laptop as the electronic device the patterned bag is designed to carry, making a reasonable inference based on the instruction's description of the bag's 'flat, rectangular shape' and its positioning as a slim travel accessory. The response also addresses the cable storage question by providing two practical options—the main suitcase or the patterned bag's internal pocket—which reflects realistic packing scenarios. The response is clear, concise, and demonstrates practical understanding of travel luggage organization. However, the response could be marginally strengthened by explicitly stating which location is more 'typically' preferred for cable storage (accessibility in the patterned bag while traveling versus storage in the main suitcase). Additionally, while 'laptop/tablet bag' is accurate, the response could have been slightly more definitive about which device is the primary target. Overall, the answer effectively fulfills the instruction's requirements."

Figure 7: Qualitative examples of compositional bias. We illustrate cases exhibiting compositional bias in MM-JudgeBias by comparing the judge outputs under unbiased and biased settings.

<b>Task Category</b>	<b>Complete List of Functional Subtask Keys</b>
<b>General</b>	Scene Recognition, Visual Composition, Spatial Organization, Object Presence/Prominence, Object Interaction, Shape/Structure Understanding, Color/Lighting Analysis, Texture/Material Perception, Pattern/Repetition Detection, Scale/Quantity Perception, Depth/Perspective Reasoning, Visual Contrast/Saliency, Contextual Consistency, Anomaly/Incongruity Detection, Aesthetic/Atmosphere Understanding.
<b>Spatial</b>	Scene Structure Understanding, Spatial Partitioning, Foreground/Background Relation, Depth and Distance Perception, Spatial Arrangement Pattern, Relative Position/Orientation, Boundary/Transition Recognition, Openness/Density Perception, Spatial Symmetry/Balance, Environmental Layout Type, Navigation/Accessibility, Lighting/Depth Correlation.
<b>Layout/Geometry</b>	Geometric Shape Recognition, Shape Composition/Decomposition, Symmetry/Proportion Analysis, Angle/Orientation Reasoning, Length/Area Estimation, Parallelism/Perpendicularity Recognition, Intersection/Connectivity Understanding, Relative Position Reasoning, Shape Transformation Awareness, 3D Projection/Perspective Inference, Geometric Pattern Recognition, Qualitative Spatial Reasoning.
<b>Text</b>	Text Presence/Localization, Text Recognition, Word-level Understanding, Sentence/Phrase Comprehension, Text-Visual Alignment, Layout/Text Position Relation, Multi-region Text Integration, Text Attribute Recognition, Contextual Reading Comprehension, Numerical/Symbolic Reasoning, Document Structure Understanding, Instructional/Functional Text Interpretation, Cross-modal Text Reasoning, Implicit Textual Inference.
<b>Chart/Plot/Diagram</b>	Chart Type Identification, Axis/Label Understanding, Data Point Recognition, Legend/Color Mapping Interpretation, Trend/Pattern Detection, Comparison/Ranking Reasoning, Value Estimation/Interpolation, Quantitative Reasoning, Multi-series Integration, Annotation-Text Relation, Visual-Semantic Alignment, Diagram Structure Understanding, Process Flow Reasoning, Infographic Compositional Understanding, Contextual Insight.
<b>Table</b>	Table Structure Recognition, Header/Title Understanding, Cell Content Reading, Row/Column Relationship, Multi-cell Data Integration, Categorical Matching, Numerical Comparison/Ranking, Aggregation/Computation, Empty/Missing Value Reasoning, Table Layout/Alignment, Cross-table Comparison, Textual-Tabular Alignment, Hierarchical Table Reasoning, Contextual Insight.
<b>Web/App/UI</b>	UI Layout Recognition, Component Type Identification, Functional Role Understanding, Text-UI Alignment, Interactive Element Detection, Navigation Flow Reasoning, Grouping/Hierarchy Recognition, Visual Emphasis/Saliency, State/Mode Interpretation, Consistency/Design Pattern Recognition, Content/Layout Relationship, Responsiveness/Adaptivity, Cross-modal Interpretation, User Intention Prediction.
<b>Factual/Commonsense</b>	Object Function Understanding, Cultural/Social Knowledge, Common Scenario Reasoning, Object-Context Relationship, Human Activity Knowledge, Physical Commonsense, Temporal Commonsense, Functional Affordance Reasoning, Cause-Effect Inference, Emotion/Intention Inference, Role/Interaction Reasoning, Object Co-occurrence Knowledge, Scene Category Knowledge, Event Plausibility Evaluation.
<b>Science</b>	Scientific Concept Recognition, Experimental Setup Interpretation, Causal Scientific Reasoning, Graph Data Interpretation, Measurement/Unit Understanding, Process/Mechanism Explanation, Hypothesis Prediction, Cross-disciplinary Science Integration.
<b>Medical</b>	Anatomical Structure Recognition, Medical Imaging Interpretation, Anomaly/Lesion Detection, Diagnostic Reasoning, Tool/Equipment Understanding, Treatment/Procedure Context, Physiological Process Explanation, Temporal Comparison/Progression.
<b>Art</b>	Artistic Style Recognition, Historical/Cultural Context, Symbolic/Thematic Interpretation, Composition/Visual Balance, Color/Lighting Analysis, Artistic Intention Reasoning, Comparative Style Reasoning, Symbol-Emotion Relationship.
<b>Causal/Logical</b>	Cause-Effect Reasoning, Conditional Scenario Reasoning, Event Sequence Understanding, Counterfactual Reasoning, Physical Logic Reasoning, Object Interaction Logic, Goal/Intention Inference, Consistency/Contradiction Detection, Analogy/Relational Reasoning, Multi-step Logical Deduction, Causal Chain Understanding, Plausibility Evaluation, Implicit Cause Detection, Conflict Resolution Reasoning.
<b>Math</b>	Number Recognition, Symbol/Expression Understanding, Geometric Relation Reasoning, Measurement/Unit Conversion, Proportional/Ratio Reasoning, Counting/Enumeration, Arithmetic Calculation, Algebraic Pattern Recognition, Graph/Coordinate Interpretation, Equation Solving Reasoning, Logical Quantitative Comparison, Word Problem Visual Translation, Spatial-Numerical Integration.
<b>Code/Symbolic</b>	Code Generation, Symbolic Expression Generation, Pseudocode Generation.
<b>Exam</b>	Problem Statement Understanding, Visual Question Interpretation, Option Comparison/Selection, Multi-modal Comprehension, Quantitative Reasoning, Logical Deduction, Diagram/Figure Analysis, Condition/Constraint Reasoning, Cross-concept Integration, Error/Misconception Identification, Sequential Problem Solving, Visual Text Parsing, Knowledge Retrieval/Application, Answer Justification.

Table 18: **Multimodal subtask specifications.** Keywords for subtasks used to ensure task diversity in multimodal query generation.

### Query-Generation Prompt Template

```
## Task Description:
You are creating a benchmark question set for evaluating a Multimodal Large Language Model (MLLM).
Given the image and contextual domain, generate exactly three distinct questions that
align with the task specifications.

## Core Requirements:
1. Cross-modal Necessity: Must require both image and text.
2. Complementary Info: Image and text must together lead to the answer.
3. No detailed image descriptions.
4. Strict Output: No explanations, output only the three questions.

## Difficulty Guidelines:
{difficulty_guidelines}

## Input Context:
Domain Description: {domain_specific_prompt}
Knowledge Guidelines: {knowledge_guidelines}

## Task Specifications:
- Task 1: {subtask_1} (Difficulty: {difficulty_1})
- Task 2: {subtask_2} (Difficulty: {difficulty_2})
- Task 3: {subtask_3} (Difficulty: {difficulty_3})

## Output Format (STRICT):
### Question1: <Question Text>
### Question2: <Question Text>
### Question3: <Question Text>
```

Figure 8: **Query-generation prompt template.** Full prompt template used for query generation.

## Query-Selection Prompt Template

```
## Task Description
You are evaluating three candidate benchmark questions (A, B, C) designed for assessing a
Multimodal Large Language Model (MLLM).
Your goal is to determine which single question (A, B, or C) is the most suitable for
evaluating multimodal understanding of the given image.
You must compare the three questions relative to one another and select only the
best-performing question based on the criteria below.
Do not include any explanations, justifications, or additional text beyond the required output
format.

## Input Questions
- Question A: {question_A}
- Question B: {question_B}
- Question C: {question_C}

## Given Information
- Image Domain type: {domain_type}
- Task type: {task_type}
- Subtask specification for A: {subtask_A}
- Subtask specification for B: {subtask_B}
- Subtask specification for C: {subtask_C}

## Evaluation Criteria
1. Bi-modal Complementarity
  - Does the question require both the image and the text to answer?
  - It must not be solvable using only the image or only the text.
2. Answerability & Unambiguity
  - Is the question clearly answerable from the given image?
  - Is the expected answer unambiguous?
3. Task Fit
  - Does the question match the given task type, domain type, and each subtask specification?
4. Content Alignment
  - Is the question semantically aligned with the image and domain?
  - Does it avoid trivial or irrelevant details?
5. Evaluation Utility
  - Does the question effectively evaluate MLLM multimodal understanding?
  - Is it neither trivial nor ill-defined?
6. Difficulty
  - Is the reasoning depth appropriate for meaningful evaluation?
  - If difficult, does it require multi-step or conceptual reasoning?
7. Instruction Clarity / Question Structure
  - Is the question linguistically clear, concise, and well-structured?
  - Does it avoid unnecessary descriptions or redundancy?

## Output Format (STRICT)
You must output exactly the following format:

### Overall Assessment:
Summary: [Brief comparative analysis of all three questions based on the evaluation
criteria]
Final Ranking: [Format: A > B > C, where the first letter is the best]
Rationale: [One-sentence explanation for the ranking based on compliance with the criteria
above]

### Selected Question: [A or B or C]
```

Figure 9: **Query-selection prompt template.** Full prompt template used for query selection.

### Text-only Query Generation Prompt Template

```
## Task Description
You are creating a question for evaluating a Large Language Model (LLM).
Your goal is to write one high-quality question that defines a task the model must perform
using ONLY the textual input.

## Core Requirements (STRICT)
1. The question must require an open-ended text response.
2. The question must be answerable, unambiguous, and not excessively long.
3. The question must be fully self-contained, providing all necessary information within
itself.
4. Do not add any system prompt before the question.
5. The question must be generated based on the domain and task specification provided below.
6. The question must align with the specified difficulty level.
7. No examples, no answers, no explanatory text.

## Domain and Task Specification
- Domain: {domain_specific_prompt}
- Task Specification (Sub-task): {task_specific_prompt}

## Difficulty Level
- Difficulty: {difficulty}

## Difficulty Guideline
{difficulty_guideline}

## Output Format (STRICT)
### Instruction: <write a single self-contained instruction here>

- Do NOT add anything else.
- Do NOT include examples.
- Do NOT require external knowledge.
- Do NOT refer to images or visual content.
```

Figure 10: **Text-only query-generation prompt template.** Full prompt template used for text-only query generation.

### Score-wise Judgment Prompt Template

```
### Task Description:
An instruction (might include an input inside it), a response to evaluate, and an image are
given.

1. Write a detailed feedback that assesses the quality of the response strictly based on how
well it follows the given instruction.
2. After writing a feedback, write a score that is an integer between 1 and 10.
3. Your response must adhere strictly to the following format:
### Feedback: (Write a feedback)
### Score: (Only output a single integer between 1 and {MAX_SCORE}, without any additional
text or explanation.)
4. Please do not generate any other opening, closing, and explanations.

### The instruction to evaluate:
{orig_instruction}

### Response to evaluate:
{orig_response}

### Feedback:
### Score:
```

Figure 11: **Score-wise judgment prompt template.** Full prompt template used for score-wise judgment.

### Pairwise Judgment Prompt Template

```
### Task Description:
An instruction (which may include an input), two responses, and an image are given.

Your task is to compare the two responses and determine which one better follows the given instruction.

1. Write a detailed feedback explaining your comparison, focusing strictly on instruction-following quality.
2. After writing the feedback, choose the better response.
3. You must select exactly one of the following options: Response A or Response B.
4. Only output your feedback and the selected response in the following format:
  ### Feedback: (Write your feedback)
  ### Selection: (Only output "A" or "B" without any additional opening, closing, and explanations.)

### The instruction to evaluate:
{orig_instruction}

### Response A:
{response_a}

### Response B:
{response_b}

### Feedback:
### Selection:
```

Figure 12: **Pairwise judgment prompt template.** Full prompt template used for pairwise judgment.

### Knowledge-Guideline Prompt Template

```
Using the original knowledge applied in the reference question, create a new question that must also rely on that knowledge to be solved.

## Reference Question: {reference_question}
```

Figure 13: **Knowledge-guideline prompt template.** Full prompt template used to guide knowledge in query generation, specifically for the *Factual/Commonsense* domain type. This enables queries to be grounded in the implicit knowledge of the source image.

### Abstention-Aware Evaluation Prompt Template

```
### Task Description:
An instruction (might include an input inside it), a response to evaluate, and an image are
given.

1. Write a detailed feedback that assesses the quality of the response strictly based on how
well it follows the given instruction.
2. After writing a feedback, write a score that is an integer between 1 and {MAX_SCORE}. If you
determine that the response cannot be properly evaluated (e.g., the instruction or response is
unclear, nonsensical, or impossible to assess), you may write "N/A" instead of a numerical
score.
3. Your response must adhere strictly to the following format: ### Feedback: (write a feedback)
### Score: (Only output a single integer between 1 and {MAX_SCORE}, or "N/A" if the response
cannot be evaluated, without any additional text or explanation.)
4. Please do not generate any other opening, closing, and explanations.

### The instruction to evaluate:
{orig_instruction}

### Response to evaluate:
{orig_response}

### Feedback:
### Score:
```

Figure 14: **Abstention-aware prompt template.** Full prompt template used for abstention-aware evaluation in the additional analysis.

### Score-Guided Evaluation Prompt Template

### Task Description:

An instruction (might include an input inside it), a response to evaluate, and an image are given.

1. Write a detailed feedback that assesses the quality of the response strictly based on how well it follows the given instruction.
2. After writing a feedback, write a score that is an integer between 1 and {MAX\_SCORE}, following the scoring guideline below:
  - 1: The response is completely irrelevant, nonsensical, or fails to address the instruction at all.
  - 2: The response barely addresses the instruction, with major errors or critical missing elements that render it largely unusable.
  - 3: The response attempts to address the instruction but contains significant errors, misunderstandings, or omissions that substantially reduce its quality.
  - 4: The response partially addresses the instruction but has notable gaps, inaccuracies, or quality issues that limit its usefulness.
  - 5: The response addresses the instruction at a basic level but lacks depth, detail, or precision, resulting in a mediocre output.
  - 6: The response reasonably addresses the instruction with some useful content, but has minor errors, omissions, or areas that could be improved.
  - 7: The response addresses the instruction well with mostly accurate and relevant content, though minor improvements could be made.
  - 8: The response is strong, accurately and thoroughly addressing the instruction with only very minor shortcomings.
  - 9: The response is excellent, comprehensively and accurately addressing the instruction with high quality and attention to detail.
  - 10: The response is perfect, flawlessly addressing every aspect of the instruction with exceptional quality and completeness.
3. Your response must adhere strictly to the following format: ### Feedback: (write a feedback) ### Score: (Only output a single integer between 1 and {MAX\_SCORE}, without any additional text or explanation.)
4. Please do not generate any other opening, closing, and explanations.

### The instruction to evaluate:  
{orig\_instruction}

### Response to evaluate:  
{orig\_response}

### Feedback:  
### Score:

Figure 15: **Score-guided prompt template.** Full prompt template used for score-guided evaluation in the additional analysis.

### Modality Constraints Evaluation Prompt Template

### Task Description:

An instruction (might include an input inside it), a response to evaluate, and an image are given.

1. You MUST carefully examine ALL of the following inputs before writing your feedback:

- The IMAGE provided (if any)
- The INSTRUCTION (question/task)
- The RESPONSE to evaluate

Do not skip or neglect any of these inputs. Your evaluation must be grounded in all available modalities.

2. Write a detailed feedback that assesses the quality of the response strictly based on how well it follows the given instruction, taking into account the image content when relevant.

3. After writing a feedback, write a score that is an integer between 1 and {MAX\_SCORE}.

4. Your response must adhere strictly to the following format: ### Feedback: (write a feedback) ### Score: (Only output a single integer between 1 and {MAX\_SCORE}, without any additional text or explanation.)

5. Please do not generate any other opening, closing, and explanations.

### The instruction to evaluate:

{orig\_instruction}

### Response to evaluate:

{orig\_response}

### Feedback:

### Score:

Figure 16: **Modality-constraints prompt template.** Full prompt template used for modality-constraints evaluation in the additional analysis.

## Modality Reasoning Evaluation Prompt Template

### ### Task Description:

An instruction (might include an input inside it), a response to evaluate, and an image are given.

To ensure a fair, rigorous, and logically consistent evaluation, you MUST follow these reasoning steps:

#### Step 1: Instruction Breakdown & Rubric Derivation (CRITICAL)

- Thoroughly and exhaustively decompose the instruction into granular, atomic, independently scorable rubric items.
- Capture every explicit and implicit requirement, condition, constraint, dependency, and expectation.
- Explicitly determine whether each requirement depends on specific modalities (e.g., image content, response content, or both).
- Do not overlook minor details.
- Each distinct requirement must be separated so it can be independently evaluated.

#### Step 2: Rubric-based Evaluation (Strict, Multi-Modal)

- Carefully examine BOTH the image (if provided) and the response.
- Evaluate the response strictly against the rubric derived in Step 1.
- Identify where the response satisfies, partially satisfies, or fails to satisfy the requirements.
- Penalize omissions, constraint violations, hallucinations, and incorrect or missing image usage when relevant.

#### Step 3: Final Synthesis & Scoring

- Synthesize the rubric-based evaluation into a coherent final assessment.
- Provide detailed feedback grounded in the rubric.
- Assign an integer score between 1 and {MAX\_SCORE} that precisely reflects the overall quality.

### ### The instruction to evaluate:

{orig\_instruction}

### ### Response to evaluate:

{orig\_response}

### ### Feedback:

(Write a detailed final feedback that reflects the reasoning process above.)

### ### Score:

(Only output a single integer between 1 and {MAX\_SCORE}, without any additional text or explanation.)

Figure 17: **Modality-reasoning prompt template.** Full prompt template used for modality-reasoning evaluation in the additional analysis.