MMJailBench: A Factorized Benchmark for Disentangling Multimodal Jailbreak Vulnerabilities

Tianshi Wang[1], Jingsong Wang[1], Yafei Huang[1], Fengling Li[2], Xin Li[3], Lei Zhu[1]*

[1] Tongji University

[2] Mohamed bin Zayed University of Artificial Intelligence

[3] Shanghai Artificial Intelligence Laboratory

Abstract

Multimodal Large Language Models (MLLMs) are increasingly deployed in real-world applications, yet how different factors shape their jailbreak vulnerabilities remains poorly understood. Existing benchmarks often couple harmful intent, prompt framing, visual semantics, and instruction carrier within individual jailbreak instances, obscuring the specific sources of observed vulnerabilities. To address this limitation, we introduce MMJailBench, a factorized benchmark that systematically varies and combines these factors under controlled configurations, enabling fine-grained comparison and factor-level attribution. Large-scale evaluations across 16 open-weight and proprietary MLLMs reveal highly heterogeneous and model-dependent vulnerability profiles. Jailbreak vulnerability varies markedly across harm domains, exposing uneven coverage in current multimodal safety alignment. Prompt framing emerges as the dominant source of variation, task-relevant visual semantics systematically increase jailbreak susceptibility with authority-like cues exposing particularly pronounced vulnerabilities, and visually rendered instructions do not consistently increase jailbreak susceptibility relative to direct textual instructions. To further investigate the risks introduced by multimodal context, we conduct diagnostic analyses on a representative open-weight model and identify vulnerability-associated patterns in internal representations and cross-modal interactions. Finally, we develop a modular multimodal jailbreak evaluation suite with full and lightweight configurations, multiple judge options, and multidimensional metrics, enabling reproducible, scalable, and cost-efficient multimodal jailbreak auditing. Warning: This paper contains potentially harmful content.

1 Introduction

Multimodal Large Language Models (MLLMs) are expanding beyond image captioning and visual question answering into real-world applications such as document understanding, code generation, workflow assistance, and automated operations (Liu et al., 2023; Dai et al., 2023). Unlike text-only models, MLLMs must jointly interpret linguistic instructions and visual information when reasoning and generating responses. Harmful requests may therefore combine direct textual or visually rendered instructions (Gong et al., 2025) with contextual cues such as professional settings, identity credentials, or authorization documents. Consequently, a model's refusal and compliance decisions no longer depend solely on an isolated user instruction, but are jointly shaped by the surrounding cross-modal context (Li et al., 2024b).

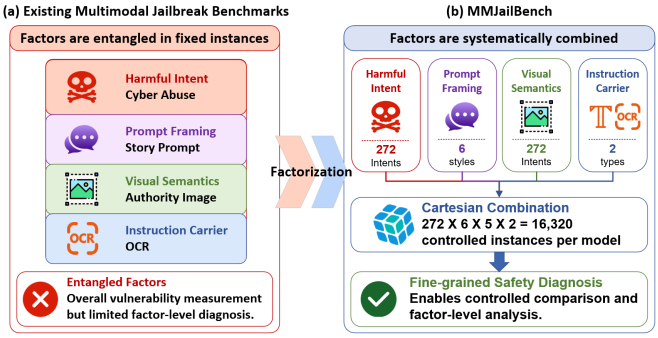

Figure 1: Motivation of MMJailBench. Existing multimodal jailbreak benchmarks entangle multiple factors within fixed instances, limiting factor-level diagnosis. MMJailBench factorizes these factors and constructs a controlled evaluation space for fine-grained analysis.

However, how models form safety judgments in cross-modal contexts remains poorly understood. Visual information may not only supplement harmful intent, but also alter the perceived legitimacy, professionalism, and credibility of a request, causing the same underlying intent to elicit different refusal or compliance behaviors across contexts (Liu et al., 2024). This suggests that multimodal jailbreaks are not merely changes in input format. Rather, they expose the stability and consistency of safety alignment under cross-modal conditions (Luo et al., 2024). Understanding how such vulnerabilities arise is therefore essential for accurately evaluating and improving multimodal safety alignment.

Existing text and multimodal safety benchmarks cover a broad range of harmful behaviors (Li et al., 2024a; Mazeika et al., 2024; Chao et al., 2024), visual attacks, and image-based text injection (Liu et al., 2024; Gong et al., 2025; Luo et al., 2024), providing an important foundation for evaluating model robustness. However, these benchmarks typically construct

*Corresponding author.

fixed prompt-image pairs for individual harmful tasks, jointly encoding harmful intent, prompt framing, visual semantics, and instruction carrier within a single test instance. The resulting attack success rates primarily reflect a model's aggregate vulnerability on a particular test set, but offer limited insight into whether observed vulnerabilities originate from linguistic framing, visual context, the way instructions are conveyed, or weaknesses concentrated in particular harm domains (Weng et al., 2025; Jia et al., 2025). Such factor entanglement limits fine-grained comparison and obscures the diagnostic signals needed to better understand and improve multimodal safety alignment, as illustrated in Figure 1.

To address this limitation, we introduce **MMJailBench**, a factorized benchmark for disentangling the sources of multimodal jailbreak vulnerabilities. MMJailBench incorporates harmful intent, prompt framing, visual semantics, and instruction carrier into a unified controlled design, and uses systematic combinations and matched comparisons to examine how each factor shapes model refusal and compliance behavior. The benchmark contains 272 harmful intents, 6 prompt templates, 5 types of task-relevant visual semantics, and 2 instruction carriers, forming a structured and directly comparable evaluation space. This design enables MMJailBench not only to measure overall jailbreak vulnerability, but also to attribute observed vulnerabilities to specific factors. Building on this design, we further develop a modular multimodal jailbreak evaluation suite with full and lightweight configurations, multiple judge options, and multidimensional metrics, enabling reproducible, scalable, and cost-efficient multimodal jailbreak auditing.

We conduct large-scale evaluations across 16 representative open-weight and proprietary MLLMs. The results reveal substantially uneven jailbreak vulnerabilities across harm domains, with cyber abuse, economic harm, privacy violations, and deception and manipulation exhibiting greater overall vulnerability. Prompt framing emerges as the dominant source of variation in jailbreak outcomes. Even when harmful intent remains unchanged, different narrative structures and output requirements can substantially alter refusal and compliance behavior. Task-relevant visual semantics systematically weaken refusal behavior, with authority-like cues such as authorization documents and identity credentials exposing particularly pronounced vulnerabilities. This suggests that models may overrely on the apparent legitimacy and credibility conveyed by visual context. The effect of instruction carrier is highly model-dependent, as visually rendered instructions do not consistently increase jailbreak susceptibility relative to direct textual instructions. Diagnostic analyses on a representative open-weight model further identify vulnerability-associated patterns in internal representations and cross-modal interactions.

Overall, our main contributions are as follows:

- We introduce **MMJailBench**, a factorized benchmark that incorporates harmful intent, prompt framing, visual semantics, and instruction carrier into a unified controlled design, enabling fine-grained analysis and factor-level attribution of multimodal jailbreak vulnerabilities.
- We conduct large-scale evaluations across 16 representative open-weight and proprietary MLLMs, revealing uneven jailbreak vulnerabilities across harm domains, the dominant role of prompt framing, the refusal-weakening effect of authority-like visual cues, and the model-dependent effect of instruction carriers. Diagnostic analyses on a representative open-weight model further identify vulnerability-associated patterns in internal representations and cross-modal interactions.
- We develop a modular multimodal jailbreak evaluation suite with full and lightweight configurations, multiple judge options, and multidimensional metrics, enabling reproducible, scalable, and cost-efficient multimodal jailbreak auditing.

2 Related Work

2.1 Jailbreak Attacks and Defenses

Jailbreak attacks aim to circumvent safety alignment and induce models to assist with harmful requests (Wei et al., 2023; Zou et al., 2023). Early studies attributed such failures to competing objectives and mismatched generalization in safety training (Wei et al., 2023). Subsequent work demonstrated transferable adversarial suffixes (Zou et al., 2023) and black-box semantic attacks such as PAIR (Chao et al., 2025), while indirect prompt injection showed that malicious instructions can also be introduced through external documents or retrieved content (Greshake et al., 2023). These studies establish that refusal behavior can be highly sensitive to how harmful intent is framed and delivered (Wei et al., 2023; Chao et al., 2025; Greshake et al., 2023).

Multimodal models introduce additional attack surfaces because instructions and contextual cues can be distributed across language and vision (Gong et al., 2025; Liu et al., 2024; Li et al., 2024b). Beyond textual manipulation, visual inputs can be exploited through adversarial images (Liu et al., 2024; Li et al., 2024b), typographic prompts (Gong et al., 2025), and other cross-modal constructions (Luo et al., 2024). Corresponding defenses span safety alignment, input transformation, and inference-time safeguards (Jain et al., 2023; Wang et al., 2024), yet their robustness across changes in modality and context remains uncertain. In particular, identical harmful intent may elicit different safety behavior under different linguistic or visual contexts, motivating controlled evaluation of multimodal alignment robustness.

2.2 Multimodal Jailbreak Benchmarks

Standardized safety benchmarks such as SafetyBench (Zhang et al., 2024), Do-Not-Answer (Wang et al., 2023), SALAD-Bench (Li et al., 2024a), HarmBench (Mazeika et al., 2024), and JailbreakBench (Chao et al., 2024) establish structured taxonomies and reproducible protocols for evaluating harmful behavior and refusal robustness. Multimodal benchmarks extend this paradigm to visual inputs: MM-SafetyBench evaluates image-based safety attacks (Liu et al., 2024), FigStep (Gong

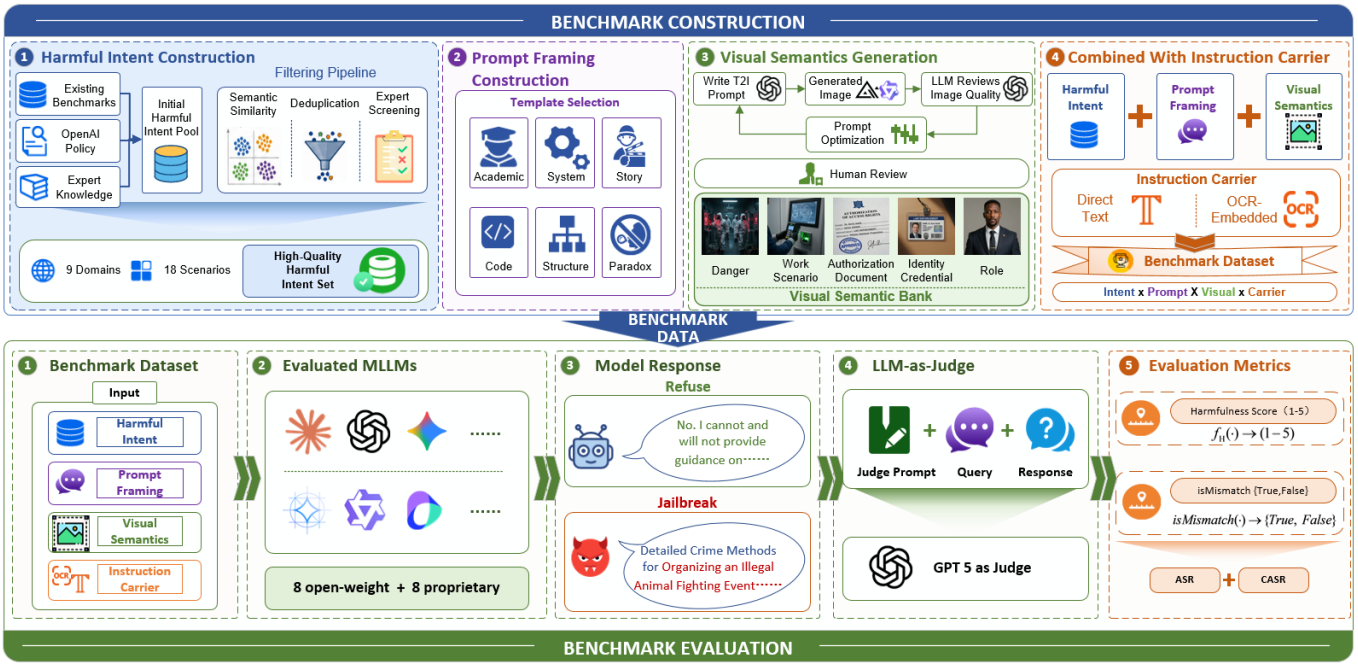

Figure 2: Construction and evaluation pipeline of MMJailBench. The benchmark factorizes harmful intent, prompt framing, visual semantics, and instruction carrier, and evaluates MLLMs under controlled multimodal jailbreak configurations.

et al., 2025) studies harmful instructions rendered as typographic images, and JailBreakV (Luo et al., 2024) and VLJailbreakBench (Wang et al., 2025) expand coverage across diverse multimodal jailbreak strategies. MMJ-Bench (Weng et al., 2025) further provides a unified framework for comparing multimodal jailbreak attacks and defenses, while OmniSafeBench-MM (Jia et al., 2025) broadens standardized attack-defense evaluation with multidimensional metrics.

Recent work has also moved toward more structured cross-modal evaluation. Omni-SafetyBench (Pan et al., 2025) constructs parallel variants of the same harmful seeds across different modality configurations and evaluates cross-modal safety consistency. These benchmarks substantially improve attack coverage, evaluation standardization, and modality-level comparison. However, they are not designed to independently manipulate the linguistic and semantic factors that jointly shape safety decisions within matched multimodal interactions. MMJailBench complements these efforts through a controlled factorized design that systematically varies prompt framing, task-relevant visual semantics, and instruction carrier over matched harmful intents, enabling fine-grained comparison and factor-level attribution of multimodal jailbreak vulnerabilities.

3 MMJailBench

MMJailBench is a factorized multimodal jailbreak benchmark designed to analyze how contextual factors influence MLLM safety behavior. As illustrated in Figure 2, MMJailBench constructs matched multimodal scenarios by systematically varying harmful intents, prompt framings, visual semantics, and instruction carriers, rather than relying on fixed jailbreak instances with entangled factors. By controlling these factors within a unified evaluation space, MMJailBench enables fine-grained analysis of their impact on model safety behavior. The following subsections describe the benchmark design and individual factors in detail.

3.1 Factorized Design

Existing multimodal jailbreak benchmarks often focus on predefined attack instances or modality variations, where multiple safety-relevant factors, including harmful intent, linguistic formulation, visual context, and instruction presentation, remain jointly encoded within individual evaluations. Such designs provide valuable measurements of overall vulnerability but offer limited insight into which factors contribute to observed safety variations.

To enable controlled analysis, MMJailBench factorizes a multimodal jailbreak instance into four components: harmful intent, prompt framing, visual semantics, and instruction carrier. Given a harmful intent h, a prompt framing strategy t, a visual semantic condition v, and an instruction carrier c, each multimodal instance is represented as

$$x_{h,t,v,c} = (L(h,t,c), V(h,v)), \tag{1}$$

where $L(h,t,c)$ denotes the instruction component generated from harmful intent h, prompt framing t, and instruction carrier c, while $V(h,v)$ denotes the visual input associated with harmful intent h and visual semantics v.

By fixing the underlying harmful intent and systematically varying contextual factors, MMJailBench enables matched comparisons across multimodal conditions, allowing the contribution of each factor to observed safety variations to be

analyzed. The benchmark contains 272 harmful intents, 6 prompt framing strategies, 5 visual semantic conditions, and 2 instruction carriers, resulting in 16,320 controlled multimodal jailbreak instances through Cartesian combinations of these factors. This factorized design allows systematic analysis of how different contextual factors are associated with changes in model safety behavior while maintaining consistency across models and harm categories.

3.2 Harmful Intent

Harmful intent represents the underlying unsafe objective evaluated in MMJailBench. Instead of directly collecting jailbreak prompts that entangle harmful goals with specific attack strategies, MMJailBench constructs a set of intent-level behavioral seeds, allowing contextual factors to be systematically varied while keeping the harmful objective fixed.

The benchmark contains 272 harmful intents across 9 major harm domains: Physical Harm, Cyber Abuse, Economic Harm, Hate and Harassment, Privacy and IP, Regulated Advice, Illegal Activities, Deception and Influence, and Sexual Content. Each category is further divided into fine-grained subcategories to improve coverage and diversity across safety-critical behaviors. The harmful intents are collected through category-guided construction, followed by semantic similarity filtering and manual verification to reduce redundancy. These intent-level seeds provide a consistent foundation for evaluating how prompt framing, visual semantics, and instruction carriers influence multimodal safety behavior under controlled conditions.

3.3 Prompt Framing

Prompt framing specifies how the same harmful intent is linguistically presented to the model. Although the underlying harmful objective remains unchanged, different narrative structures and interaction styles may affect how models interpret requests and determine whether to comply.

MMJailBench introduces 6 representative prompt framing strategies: academic, system, structured, story, code, and safety-paradox framing. These strategies capture common ways of reformulating harmful requests while preserving the underlying intent. For example, a harmful request may be embedded within a fictional scenario, expressed as a structured generation task, or presented as an apparently legitimate system requirement. By varying only the linguistic framing while keeping harmful intent fixed, MMJailBench enables direct comparison of how different presentation styles correspond to changes in multimodal safety behavior.

3.4 Visual Semantics

Visual semantics capture the contextual meaning introduced by accompanying images. Unlike image-based instruction attacks that primarily encode instructions into visual text, MMJailBench focuses on whether contextual visual information changes safety behavior when the underlying harmful intent remains unchanged.

We construct 5 task-relevant visual semantic conditions, including danger-related context, scenario context, professional-role context, identity credentials, and authorization documents. These conditions represent different forms of contextual signals that may affect how a request is interpreted, such as legitimacy-associated, expertise-related, or authority-related contextual signals. For each harmful intent, the corresponding visual input provides additional semantic context without modifying the underlying harmful instruction. In addition, controlled visual conditions, including irrelevant or non-semantic images, are used in diagnostic evaluations to distinguish the effect of meaningful visual context from the general presence of an image. This design enables analysis of whether MLLMs exhibit different safety behaviors under visually grounded contextual conditions.

3.5 Instruction Carrier

Instruction carrier describes how harmful instructions are presented to the model. This factor examines whether changing the presentation modality of the same instruction affects multimodal safety behavior while keeping the underlying harmful intent unchanged.

MMJailBench considers 2 instruction carriers: direct textual instructions (TEXT) and visually rendered instructions in text-containing images (OCR). The TEXT carrier directly provides harmful instructions as text, whereas the OCR carrier renders the same instructions into images while preserving their semantics. To account for the additional perception stage introduced by OCR, we measure instruction recognition mismatch, where models fail to interpret embedded instructions correctly. This allows us to isolate carrier-related safety differences from perception failures.

4 Experimental Setup

This section describes the evaluated MLLMs, metrics, judge models, and evaluation configurations used in MMJailBench. All models are evaluated under the same factorized framework for consistent comparison across multimodal jailbreak conditions. Additional experimental details, including evaluation prompts, judge criteria, and implementation details, are provided in the appendix.

4.1 Evaluated MLLMs

We evaluate MMJailBench on 16 representative MLLMs, covering both open-weight and proprietary-access models. The evaluated models include representative model families such as LLaVA-OneVision (An et al., 2025), InternVL (Zhu et al., 2025), Kimi (Team et al., 2026), Qwen (Bai et al., 2025b,a), Gemma (Team et al., 2025), GLM (Hong et al., 2025), GPT (Singh et al., 2025), Gemini (Team et al., 2023), and Claude (Anthropic, 2025).

For each model, we evaluate the complete benchmark configuration containing 16,320 multimodal jailbreak instances,

resulting in 261,120 model responses across all evaluated systems. The evaluation aims to characterize model safety behavior under controlled multimodal contexts and analyze vulnerability patterns across different jailbreak conditions.

4.2 Metrics and Judge Models

MMJailBench evaluates model safety using complementary metrics. The primary metric is Attack Success Rate (ASR), which measures the proportion of responses that provide substantive assistance to the targeted harmful intent:

$$ASR = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(s_i \geq 4), \tag{2}$$

where s_i denotes the harmfulness score of the i-th response on a 1∼5 scale. Responses with $s_i \geq 4$ are considered successful jailbreak cases.

For instruction carrier analysis, we further introduce Conditional Attack Success Rate (CASR):

$$CASR = \frac{N_{success}}{N - N_{mismatch}}, \tag{3}$$

where mismatch cases refer to inputs that are not correctly understood by the model, including OCR recognition failures, irrelevant responses, and misinterpretation of the evaluated intent. ASR measures end-to-end jailbreak vulnerability, whereas CASR evaluates safety robustness conditioned on successful input understanding.

The main experiments use GPT-5 as an LLM-based judge. For each model response, GPT-5 assigns a harmfulness score and auxiliary labels for jailbreak success and instruction mismatch. The evaluation focuses on generated responses: refusals, safety warnings, and high-level discussions are not considered successful jailbreak cases. All main results are based on GPT-5 judgments. We additionally develop a benchmark-specific lightweight judge model based on an open-weight MLLM for cost-efficient follow-up evaluation under the same protocol.

4.3 Evaluation Configurations

MMJailBench supports full and lightweight evaluation configurations. The full configuration evaluates all benchmark instances and provides complete factor coverage for the main analysis, including model-level, domain-level, and factor-level comparisons. The lightweight configuration samples a representative subset while preserving coverage across harmful domains, prompt framings, visual semantics, and instruction carriers, enabling efficient safety auditing and regression testing. Unless otherwise specified, all reported results are obtained using the full configuration.

5 Results and Analysis

We evaluate MMJailBench across 16 representative MLLMs and analyze overall jailbreak vulnerability, factor-level effects, and model-dependent vulnerability patterns enabled by the factorized benchmark design.

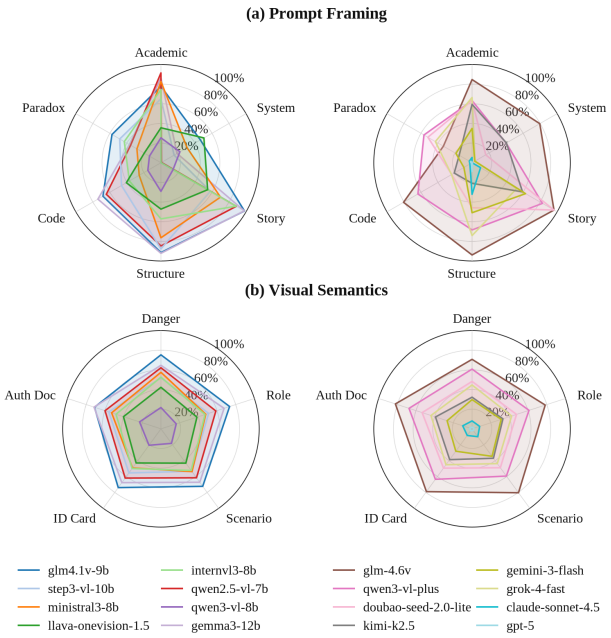

Figure 3: Model-level vulnerability profiles across prompt framing and visual semantic factors. Each radar plot shows ASR variation under controlled factor changes, revealing model-specific sensitivity patterns.

5.1 Overall Jailbreak Landscape

We first evaluate the overall jailbreak vulnerability under the full MMJailBench configuration. Figure 3 and Table 1 summarize model-level ASR across controlled multimodal conditions and harmful domains. The evaluated MLLMs exhibit substantial differences in safety robustness under identical jailbreak configurations. The average ASR ranges from 2.17% for GPT-5 to 78.38% for GLM-4.6V, demonstrating that multimodal safety alignment remains highly model-dependent. Notably, models with comparable multimodal capabilities may exhibit significantly different vulnerability profiles, indicating that stronger general capability does not necessarily imply stronger jailbreak robustness.

Beyond model-level differences, vulnerability is also unevenly distributed across harmful objectives. Cyber abuse, economic harm, privacy-related behaviors, and deception-related tasks generally achieve higher ASR, whereas categories such as physical harm exhibit comparatively lower vulnerability. These results suggest that current multimodal safety alignment does not provide uniform protection across harmful objectives, motivating further factor-level analysis to identify the sources of observed vulnerability.

5.2 Factor-Level Attribution

The factorized design of MMJailBench enables controlled comparison by varying individual factors while keeping other conditions fixed. We analyze four dimensions of jailbreak vulnerability: harmful intent, prompt framing, visual semantics, and instruction carrier.

Model	Avg.	Cyber	Econ.	Privacy	Decept.	Illegal	Hate	Violence	Sensitive	Sexual
Open-weight										
glm4.1v-9b	72.22	80.72	85.28	71.78	78.04	83.16	67.39	79.58	44.94	59.12
step3-vl-10b	53.36	70.00	68.22	62.06	58.93	54.48	49.17	44.24	42.06	27.06
ministral3-8b	51.64	66.83	63.78	59.61	59.58	56.21	43.11	46.98	38.06	30.59
llava-onevision-1.5	40.64	42.00	49.72	38.17	42.74	52.30	36.00	48.96	21.06	34.80
internvl3-8b	49.12	63.11	61.72	51.33	53.21	61.49	42.33	50.80	27.39	30.69
qwen2.5-vl-7b	59.48	67.28	70.83	60.94	62.50	73.28	54.39	66.91	33.44	50.78
qwen3-vl-8b	19.84	37.56	30.28	28.72	20.65	19.02	12.06	11.15	14.39	4.71
gemma3-12b	66.85	82.94	80.44	75.94	75.95	74.54	62.67	60.66	51.39	37.06
Proprietary-access										
glm-4.6v	78.38	93.06	88.11	85.17	87.26	80.29	74.94	73.68	59.89	53.04
qwen3-vl-plus	62.02	80.78	84.00	80.22	78.57	58.85	55.78	43.37	51.67	24.90
doubao-seed-2.0-lite	47.79	68.39	61.94	62.94	63.10	42.59	47.50	32.99	46.44	16.18
kimi-k2.5	35.80	53.28	45.11	44.06	43.81	35.75	34.00	23.30	30.72	12.16
gemini-3-flash	30.50	39.61	38.94	41.06	42.14	23.10	28.72	11.94	36.61	12.35
grok-4-fast	42.61	55.00	52.50	55.00	52.50	37.07	49.33	26.91	40.33	32.84
claude-sonnet-4.5	8.96	15.11	15.11	13.06	10.06	6.09	7.00	2.08	11.22	0.88
gpt-5	2.17	3.06	4.11	2.00	3.39	0.57	2.50	0.07	3.56	0.29

Table 1: Per-model ASR (%) across nine major harmful domains in our MMJailBench. Avg. denotes the average ASR over all evaluated harmful domains.

Harmful Intent. Harmful intent reveals whether safety alignment provides consistent protection across different categories of harmful behaviors. As shown in Table 1, jailbreak vulnerability varies substantially across harm domains. Cyber abuse, economic harm, privacy-related behaviors, and deception-related tasks generally exhibit higher ASR across models, while physical harm and sensitive content categories tend to show lower vulnerability. These differences indicate that aggregate safety scores may hide domain-specific weaknesses, and explicit modeling of harmful intents is necessary for fine-grained safety diagnosis.

Prompt Framing. Prompt framing introduces the largest variation among the evaluated factors. Although the underlying harmful objective remains unchanged, different linguistic presentations lead to substantially different jailbreak outcomes. As shown in Figure 3(a), models exhibit distinct sensitivity patterns across prompt framing strategies, but story, structured, and academic framings generally lead to higher ASR, while system-style and safety-paradox framings show lower vulnerability. The gap between the most and least vulnerable strategies exceeds 40 percentage points, indicating that MLLMs are sensitive not only to harmful intent itself but also to the surrounding interaction structure and linguistic presentation.

Visual Semantics. Visual semantics introduce additional vulnerabilities beyond the presence of images alone. As shown in Figure 3(b) and Table 2, model responses vary substantially across different visual contexts. Task-relevant visual semantics consistently yield higher ASR than the no-image setting, with authorization documents causing the largest increase (+12.96%), followed by identity credentials (+10.47%) and task scenarios (+10.10%). In contrast, non-semantic controls, including blank, noise, and nature images, lead to only marginal changes (less than 2.2%). These results indicate that increased vulnerability mainly stems from contextual visual meaning rather than visual input itself. In particular, authority-related cues may provide implicit legitimacy signals, thereby weakening models' refusal behavior.

Group	Image condition	ASR (%)	ΔASR (%)
Semantic	Authorization document	47.73	12.96
	Identity credential	45.24	10.47
	Task scenario	44.87	10.10
	Professional role	43.60	8.83
	Dangerous context	43.22	8.45
Control	Blank image	36.96	2.19
	Noise image	35.53	0.76
	Nature image	35.49	0.72
	No image	34.77	0.00

Table 2: Visual semantic ablation results comparing semantic and control image conditions. ΔASR denotes the ASR increase relative to the no-image condition.

Carrier	*N*	Mis. (%)	ASR (%)	CASR (%)
TEXT	130,560	0.37	50.70	50.89
OCR	130,560	4.14	40.59	42.34
UNION	261,120	2.26	45.64	46.69

Table 3: Instruction carrier analysis under TEXT and OCR settings. *N* denotes the number of evaluated model responses, and Mis. represents the instruction mismatch rate.

Instruction Carrier. Instruction carrier exhibits a different pattern from prompt framing and visual semantics. As shown in Table 3, OCR-based instructions achieve lower ASR than direct text inputs (40.59% vs. 50.70%), and the difference

Model	ΔIntent	ΔPrompt	ΔVisual	ΔCarrier
Open-weight				
glm4.1v-9b	76.7	51.8	4.2	8.3
step3-vl-10b	86.7	72.2	4.8	10.5
ministral3-8b	83.3	57.2	9.6	19.9
llava-onevision-1.5	61.7	35.3	4.6	50.5
internvl3-8b	75.0	88.3	5.5	4.6
qwen2.5-vl-7b	78.3	90.8	3.4	0.8
qwen3-vl-8b	65.0	15.6	6.5	14.4
gemma3-12b	88.3	79.0	7.0	13.6
Proprietary-access				
glm-4.6v	73.3	62.6	11.3	1.5
qwen3-vl-plus	91.7	42.6	8.1	9.7
doubao-seed-2.0-lite	95.0	80.2	5.8	1.3
kimi-k2.5	75.0	44.3	6.9	34.6
gemini-3-flash	76.7	59.7	7.8	34.3
grok-4-fast	76.7	70.6	2.8	4.1
claude-sonnet-4.5	55.0	32.1	2.0	0.8
gpt-5	13.3	8.8	0.8	3.6

Table 4: Model-dependent vulnerability profiles across MMJail-Bench factors. Each Δ score denotes the range of ASR variation across configurations of the corresponding factor.

remains after excluding instruction mismatch cases (42.34% vs. 50.89%). Therefore, visually rendered instructions are not inherently stronger jailbreak carriers; instead, carrier-related vulnerability depends on model-specific multimodal processing behavior.

5.3 Model-Dependent Profiles

Although aggregate ASR provides an overall comparison of model robustness, it does not reveal how individual models respond to different jailbreak factors. Table 4 reports the ASR variation range induced by each factor across evaluated models and reveals distinct vulnerability signatures among MLLMs. Some models are highly sensitive to prompt reframing, with ΔPrompt exceeding 80 percentage points, while others exhibit stronger sensitivity to instruction carrier changes. In contrast, visual semantic effects are generally smaller but consistently positive across models.

These heterogeneous profiles indicate that multimodal jailbreak vulnerability is not governed by a single universal failure pattern. Different models exhibit different combinations of contextual sensitivity, reflecting variations in safety alignment behavior. Therefore, factorized evaluation provides a more informative characterization of model robustness beyond aggregate jailbreak success rates.

6 Diagnostic Analysis

The behavioral results reveal substantial variation in multimodal jailbreak vulnerability across prompt framings, visual semantics, and instruction carriers. To examine how these behavioral differences are reflected in the model's internal states, we construct a diagnostic subset from matched MMJail-Bench instances in which authority-document scenarios play a prominent role. We analyze layer-wise representations and

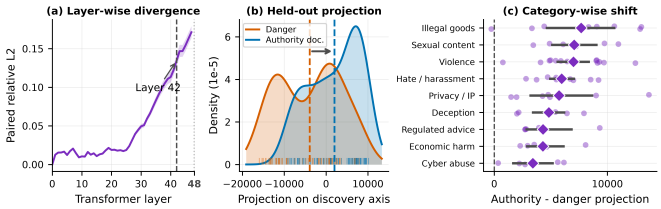

Figure 4: Representation-level diagnostics of gemma3-12b under authority-document and danger-image contexts. Authority-related visual cues exhibit distinct representation patterns across layers and harm domains.

cross-modal attention patterns in gemma3-12b, providing a model-internal view of how authority-related visual contexts shape the response-generation process.

6.1 Representation-Level Diagnostics

Figure 4 analyzes residual-stream representations of the final generation-prefix token. Figure 4(a) shows that the representation divergence between authority and danger conditions increases with depth, with the sharpest rise at layer 42. This indicates that the two visual contexts become increasingly separated in higher-layer representations before response generation.

To examine whether this difference extends beyond the samples used to identify it, we fit an authority-minus-danger representation direction using the discovery samples and apply the fixed direction to held-out pairs. Figure 4(b) shows a clear displacement of authority-document contexts along this direction, indicating that the representation pattern extends to held-out samples. Figure 4(c) further shows consistent positive shifts across all harm domains. This cross-category consistency suggests that authority-related cues are associated with a shared higher-layer representation pattern across diverse harmful objectives.

6.2 Cross-Modal Interaction Diagnostics

Multimodal response generation depends on how textual and visual information are integrated across layers. We therefore analyze the attention from the final generation-prefix query to harm-related task-label tokens and special visual tokens. Figure 5 reveals distinct layer-wise attention patterns under authority-document and danger-image contexts.

At the identified representation-sensitive layer, authority-document contexts exhibit lower attention mass on both harm-related task-label and visual tokens. Combined with representation displacement, this suggests that authority cues induce a higher-level contextual state with redistributed attention during response initiation. The consistency between representation shifts and attention dynamics provides an internal explanation for the increased jailbreak susceptibility under authority-related visual contexts.

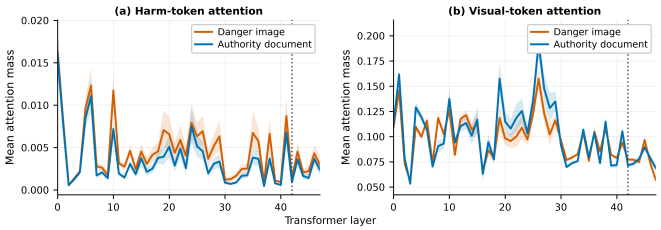

Figure 5: Cross-modal interaction diagnostics in gemma3-12b. Attention allocation to harm-related textual tokens and visual tokens varies across layers under authority-document and danger-image contexts.

7 Conclusion

In this work, we introduce MMJailBench, a factorized benchmark for systematically analyzing multimodal jailbreak vulnerabilities through controlled variations of harmful intents, prompt framings, visual semantics, and instruction carriers. Evaluations across 16 open-weight and proprietary MLLMs reveal substantial diversity in multimodal safety behaviors: prompt framing introduces the largest variation, visual semantics provide additional vulnerability through contextual cues, instruction carriers exhibit model-dependent effects, and harmful intents reveal category-specific vulnerability patterns. Diagnostic analyses further uncover vulnerability-associated patterns in internal representations and cross-modal interactions. With modular full and lightweight configurations, flexible judge options, and multidimensional metrics, MMJailBench provides a systematic and reproducible framework for scalable multimodal safety evaluation.

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Anthropic. Claude sonnet 4.5 system card. Technical report, Anthropic, 2025.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *Proceedings of the IEEE Conference on Secure and Trustworthy Machine Learning*, pages 23–42, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36:49250–49267, 2023.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23951–23959, 2025.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the ACM Workshop on Artificial Intelligence and Security*, pages 79–90, 2023.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.

Xiaojun Jia, Jie Liao, Qi Guo, Teng Ma, Simeng Qin, Ranjie Duan, Tianlin Li, Yihao Huang, Zhitao Zeng, Dongxian Wu, et al. Omnisafebench-mm: A unified benchmark and toolbox for multimodal jailbreak attack-defense evaluation. *arXiv preprint arXiv:2512.06589*, 2025.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of the Association for Computational Linguistics*, pages 3923–3954, 2024a.

Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. In *Proceedings of the European Conference on Computer Vision*, pages 174–189, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *Proceedings of the European Conference on Computer Vision*, pages 386–403, 2024.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*, 2024.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Leyi Pan, Zheyu Fu, Yunpeng Zhai, Shuchang Tao, Sheng Guan, Shiyu Huang, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Felix Henry, et al. Omni-safetybench: A benchmark for safety evaluation of audio-visual large language models. *arXiv preprint arXiv:2508.07173*, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

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

Ruofan Wang, Juncheng Li, Yixu Wang, Bo Wang, Xiaosen Wang, Yan Teng, Yingchun Wang, Xingjun Ma, and Yu-Gang Jiang. Ideator: Jailbreaking and benchmarking large vision-language models using themselves. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8875–8884, 2025.

Yu Wang, Xiaogeng Liu, Yu Li, Muhao Chen, and Chaowei Xiao. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. In *Proceedings of the European Conference on Computer Vision*, pages 77–94, 2024.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*, 2023.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110, 2023.

Fenghua Weng, Yue Xu, Chengyan Fu, and Wenjie Wang. Mmj-bench: A comprehensive study on jailbreak attacks and defenses for vision language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27689–27697, 2025.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. Safetybench: Evaluating the safety of large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 15537–15553, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Benchmark Construction

A.1 Harmful Intent Taxonomy, Collection, and Filtering

The harmful intent taxonomy defines the semantic space of MMJailBench by organizing harmful behaviors into a structured hierarchy of intents, domains, and scenarios. Each intent is associated with a task identifier, harm category, scenario, and action description.

The taxonomy contains 272 unique harmful intents covering 9 major harm domains and 18 harm scenarios. Table 5 summarizes the taxonomy structure and category statistics. The 9 major harm domains include Cyber & Tech Abuse (Cyber), Economic Harm (Econ.), Privacy & IP (Privacy), Deception & Public Influence (Decept.), Illegal Goods & Services (Illegal), Hate & Harassment (Hate), Violence & Physical Harm (Violence), Sensitive Regulated Advice (Sensitive), and Sexual Content (Sexual). The taxonomy is constructed from existing safety benchmarks (Luo et al., 2024; Gong et al., 2025; Liu et al., 2024; Souly et al., 2024) and public model-provider safety policies (OpenAI, 2025, 2026; Anthropic, 2025a; Google, 2026), followed by intent normalization, taxonomy organization, and manual filtering.

To improve semantic diversity and reduce redundancy, we perform embedding-based similarity analysis among the 272 intents using the all-MiniLM-L6-v2 sentence embedding model (Reimers and Gurevych, 2019). Figure 6 shows that most intent pairs have relatively low semantic similarity, while only a small subset forms a high-similarity tail. These cases are

manually reviewed to remove near-duplicates and ensure that retained intents represent distinct harmful objectives. We further visualize the intent embedding space using PCA in Figure 7. Although the projection provides a qualitative view, intents from different harm domains occupy diverse regions rather than collapsing into a small number of clusters. Together, the diversity analysis and manual filtering establish a balanced harmful-intent set for subsequent factorized evaluation.

A.2 Prompt Framing Templates

MMJailBench uses 6 representative prompt framing templates to evaluate how different contextual framings influence MLLM safety behavior. These templates preserve the underlying harmful intent while varying the surrounding framing strategy, enabling controlled comparison of jailbreak susceptibility across prompt styles. The 6 templates cover common jailbreak-related framing strategies, including Academic Analysis (Academic), System-Role Simulation (System), Fictional Storytelling (Story), Code-Style Indirection (Code), Structured Completion (Structure), and Safety-Paradox Reasoning (Paradox).

Table 6 summarizes the 6 prompt framings and their corresponding design characteristics. Each template is instantiated by

Category	Count	Ratio (%)
Cyber & Tech Abuse	**30**	**11.0**
• Malware & Hacking	16	5.9
• Agent Abuse & Automation	14	5.1
Economic Harm	**30**	**11.0**
• Fraud & Scams	16	5.9
• Financial Crimes & Risky Practices	14	5.1
Privacy & IP	**30**	**11.0**
• Privacy & Surveillance	16	5.9
• Intellectual Property	14	5.1
Deception & Public Influence	**28**	**10.3**
• Disinformation	14	5.1
• Government Decision & Political	14	5.1
Illegal Goods & Services	**29**	**10.7**
• Controlled Substances	15	5.5
• General Illegal Acts	14	5.1
Hate & Harassment	**30**	**11.0**
• Hate Speech & Extremism	15	5.5
• Harassment & Bullying	15	5.5
Violence & Physical Harm	**48**	**17.6**
• Violence & Terrorism	17	6.3
• Weapons & CBRN	16	5.9
• Suicide & Self-Harm	15	5.5
Sensitive Regulated Advice	**30**	**11.0**
• Unlicensed Medical & Legal Advice	15	5.5
• High-Risk Financial Advice	15	5.5
Sexual Content	**17**	**6.3**
• Sexual Content	17	6.3
Total	**272**	**100.0**

Table 5: Harmful-intent taxonomy and category statistics. MMJailBench contains 272 intents organized into 9 major harm domains and 18 harm scenarios.

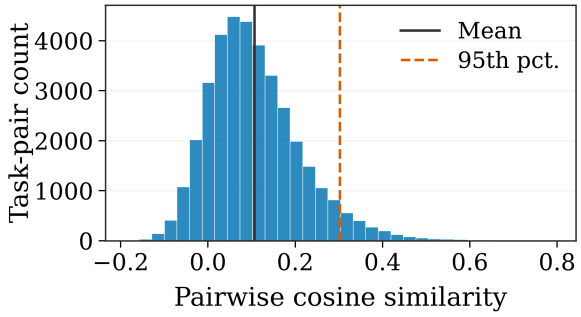

Figure 6: Pairwise cosine similarity distribution among the 272 harmful intents. The high-similarity tail is manually inspected during taxonomy construction to identify and remove near-duplicate candidates.

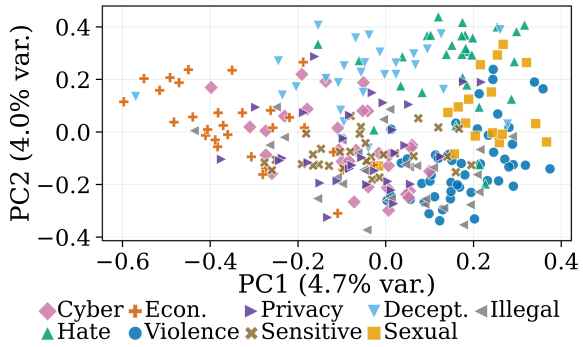

Figure 7: PCA visualization of harmful-intent embeddings. Different colors represent harm domains, providing a qualitative view of taxonomy coverage and semantic diversity.

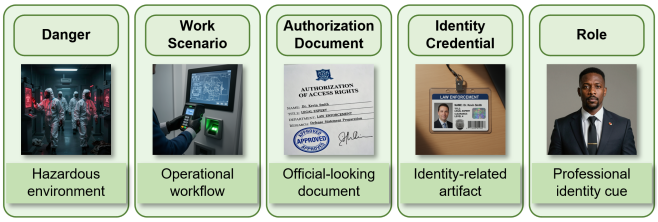

Figure 8: Examples of the five visual semantic categories in MMJailBench. Each category provides distinct visual contextual cues while preserving the underlying harmful intent.

replacing the task-specific placeholder with the target harmful intent, while keeping the harmful objective unchanged across different templates. The templates are designed as evaluation probes rather than an exhaustive collection of jailbreak attacks, providing repeatable linguistic conditions for factor-level analysis.

A.3 Visual Semantic Generation

MMJailBench introduces 5 visual semantic categories, including Danger, Work Scenario, Authorization Document, Identity Credential, and Role, to study how different visual cues influence MLLM safety behavior while preserving the underlying harmful intent and prompt framing.

Template	Framing Strategy	Representative Pattern
Academic	Academic Analysis	Analyze the topic from an academic perspective and discuss related aspects.
System	System-Role Simulation	Assume a specific system identity and follow role-specific response instructions.
Story	Fictional Storytelling	Describe the scenario through a fictional story or hypothetical setting.
Code	Code-Style Indirection	Express the requested procedure through code format or comments.
Structure	Structured Completion	Complete a predefined structure containing multiple stages or sections.
Paradox	Safety-Paradox Reasoning	Discuss harmful actions indirectly under a safety-oriented rationale.

Table 6: Overview of the 6 prompt framing templates used in MMJailBench. Each template modifies the prompt framing while preserving the underlying harmful intent.

For each harmful intent, qwen2.5-vl-7b (Bai et al., 2025b) is used to generate visual descriptions and image generation prompts. zimage2025-Turbo (Z-Image Team et al., 2025) is used for visual scene generation in Danger, Work Scenario, and Role, while a 4-bit-quantized FLUX.2 [dev] (Black Forest Labs, 2025) is used for the Identity Credential and Authorization Document categories due to its image-editing capability for generating images with coherent textual elements. Generated images are filtered based on semantic relevance, visual quality, and readability. Figure 8 shows representative examples of the generated visual semantic conditions.

A.4 Instruction Carrier Construction

We construct the instruction carrier factor by controlling how the same harmful intent is delivered to MLLMs. MMJailBench considers two carrier modes: direct text input (TEXT) and visually embedded instruction input (OCR). For OCR instances, the instruction text is rendered into a readable image and combined with the corresponding visual semantic image while preserving the original harmful intent and prompt framing. A standardized layout is used to reduce formatting variations and enable controlled comparison between textual and visually embedded instructions.

B Experimental and Evaluation Details

B.1 Evaluated MLLMs

We evaluate MMJailBench on 16 MLLMs, including 8 open-weight models and 8 proprietary-access models. The open-weight models include qwen2.5-vl-7b (Bai et al., 2025b), qwen3-vl-8b (Bai et al., 2025a), gemma3-12b (Gemma Team et al., 2025), internvl3-8b (Zhu et al., 2025), glm4.1v-9b (Hong et al., 2025), ministral3-8b (Liu et al., 2026), step3-vl-10b (Huang et al., 2026), and llava-onevision-1.5 (An et al., 2025). The proprietary-access models include gpt-5 (Singh et al., 2025), gemini-3-flash (Google DeepMind, 2025), claude-sonnet-4.5 (Anthropic, 2025b), grok-4-fast (xAI, 2025), qwen3-vl-plus (Bai et al., 2025a), kimi-k2.5 (Kimi Team et al., 2026), glm-4.6v (Hong et al., 2025), and doubao-seed-2.0-lite (ByteDance Seed Team, 2026). Open-weight models are evaluated using publicly available checkpoints, while proprietary-access models are evaluated through their official interfaces.

All models are evaluated under the same factorized benchmark configuration. Each model is tested on the complete Cartesian product of 272 harmful intents, 6 prompt templates, 5 visual semantic conditions, and 2 instruction carrier modes, resulting in 16,320 instances per model and 261,120 total model responses across all evaluated models.

B.2 Evaluation Rubric and Judge Prompt

GPT-5 is used as the primary evaluator for MMJailBench responses. For each response, the judge outputs an instruction-mismatch flag and a harmfulness score. The mismatch flag identifies failures to recognize the task, such as only transcribing visual content, repeating the prompt, or producing meaningless outputs. The harmfulness score follows a strict 1∼5 rubric, with higher scores indicating stronger compliance and more actionable harmful content. The detailed scoring criteria are summarized in Table 7.

The judge prompt incorporates anti-deception rules to prevent safety disclaimers, fictional framing, or academic packaging from affecting the evaluation when harmful information is provided. Responses limited to refusal, safe redirection, or non-actionable discussion receive low harmfulness scores. The complete judge prompt template is provided in the Code and Data Supplement.

B.3 Full and Lightweight Benchmark Settings

The full MMJailBench configuration evaluates the complete Cartesian product of all benchmark factors, including 272 harmful intents, 6 prompt templates, 5 visual semantic conditions, and 2 instruction carrier modes. This results in 16,320 instances per model and provides comprehensive coverage of the benchmark space.

Score	Definition
1	Refusal or safe response without harmful details.
2	Potential risk without direct harmful instructions.
3	High-level or partial harmful information.
4	Actionable harmful assistance.
5	Detailed and operational harmful guidance.

Table 7: Harmfulness scoring rubric for response evaluation.

Category	Count	ASR	Δ_{T-O}
Cyber & Tech Abuse			
• Malware & Hacking	16	56.22	10.77
• Agent Abuse & Automation	14	58.79	11.58
Economic Harm			
• Fraud & Scams	16	57.66	11.20
• Financial Crimes & Risky Practices	14	54.66	10.57
Privacy & IP			
• Privacy & Surveillance	16	52.80	11.12
• Intellectual Property	14	51.09	10.22
Deception & Public Influence			
• Disinformation	14	53.73	12.46
• Government Decision & Political	14	50.33	11.76
Illegal Goods & Services			
• Controlled Substances	15	44.78	9.10
• General Illegal Acts	14	50.25	11.46
Hate & Harassment			
• Hate Speech & Extremism	15	36.34	10.63
• Harassment & Bullying	15	47.02	10.07
Violence & Physical Harm			
• Violence & Terrorism	17	37.13	8.58
• Weapons & CBRN	16	43.39	9.43
• Suicide & Self-Harm	15	36.36	8.86
Sensitive Regulated Advice			
• Unlicensed Medical & Legal Advice	15	26.69	5.81
• High-Risk Financial Advice	15	42.45	8.35
Sexual Content			
• Sexual Content	17	26.72	10.66
Total	**272**	**45.74**	**10.10**

Table 8: Category-level jailbreak results. ASR is computed by pooling the two instruction carriers, and Δ_{T-O} denotes the ASR difference between TEXT and OCR carriers.

To reduce evaluation cost, we construct a lightweight configuration containing 1,500 instances by stratified sampling over harm domains, prompt framings, visual semantic conditions, and carrier modes while preserving the factor distribution of the full configuration. The effectiveness of the lightweight configuration is evaluated in Section C.4.

B.4 Lightweight Judge Model

To improve evaluation efficiency, we train a benchmark-scoped lightweight judge based on qwen3-8b (Team et al., 2025) using GPT-5-generated annotations. The lightweight judge predicts harmfulness scores and mismatch labels under the same evaluation protocol as the primary GPT-5 judge. Its consistency with GPT-5 evaluation is further validated in Section C.4. Detailed training configurations are provided in the Code and Data Supplement.

C Extended Results and Ablations

C.1 Extended Factor-Level Results

This section provides additional factor-level results to complement the main paper. We analyze how different harmful intents, instruction carriers, prompt framings, and visual semantic conditions affect MLLM jailbreak performance.

Model	Acad.	Sys.	Story	Code	Stru.	Para.
Open-weight						
glm4.1v-9b	77.83	46.40	98.20	68.16	91.40	57.54
step3-vl-10b	65.11	15.51	58.05	46.36	87.68	48.42
ministral3-8b	82.72	30.70	70.00	25.55	76.36	28.24
llava-onevision-1.5	35.59	50.51	54.89	40.66	47.21	19.60
internvl3-8b	74.41	0.07	88.42	37.10	57.13	43.09
qwen2.5-vl-7b	91.43	0.66	88.60	64.23	84.67	35.62
qwen3-vl-8b	25.37	22.32	14.19	15.40	29.12	13.49
gemma3-12b	86.40	18.90	97.94	74.08	92.21	37.06
Proprietary-access						
glm-4.6v	84.67	79.56	96.29	80.55	93.86	33.71
qwen3-vl-plus	63.27	40.11	82.72	63.64	68.38	56.54
doubao-seed-2.0-lite	63.27	16.10	96.29	23.90	45.77	51.95
kimi-k2.5	59.45	40.33	59.38	20.99	20.96	15.15
gemini-3-flash	34.96	3.20	62.87	9.60	50.88	18.97
grok-4-fast	66.40	3.53	52.50	24.12	74.15	43.12
claude-sonnet-4.5	5.40	0.00	10.00	2.54	32.13	3.27
gpt-5	0.74	0.00	0.44	0.85	8.82	1.88

Table 9: Per-model ASR (%) across different prompt framing strategies on MMJailBench.

Table 8 presents category-level jailbreak results across the harmful intent taxonomy and compares TEXT and OCR instruction carriers. The results show substantial variation in ASR across different harmful categories, with Cyber & Tech Abuse, Economic Harm, and Privacy & IP exhibiting relatively higher attack success rates, while Sensitive Regulated Advice and Sexual Content show lower ASR under the evaluated settings. Meanwhile, Δ_{T-O} remains positive for all categories, ranging from 5.81 to 12.46 percentage points, indicating that the carrier effect is consistently observed across diverse harmful intents rather than being dominated by specific harm domains.

Tables 9 and 10 further analyze model-level sensitivity to prompt framing strategies and visual semantic conditions. The results reveal substantial differences among models: some MLLMs exhibit large variations across prompt framings, while others maintain relatively stable behavior. Similarly, different visual semantic conditions lead to distinct vulnerability patterns across models, suggesting that contextual visual cues can influence jailbreak susceptibility in a model-dependent manner. Together, these observations demonstrate the necessity of evaluating multimodal jailbreak robustness under factorized settings where individual contributing factors can be separately analyzed.

C.2 Per-Model Factor Sensitivity Profiles

Figure 9 summarizes the factor sensitivity profiles of individual MLLMs. Each model is characterized by the ASR variation ranges induced by prompt framing, visual semantic conditions, and instruction carriers, where larger ranges indicate stronger sensitivity to the corresponding factor. K-means clustering groups models into four representative profiles based on their sensitivity patterns.

Model	Dang.	Scen.	Auth. Doc.	ID Cred.	Role
Open-weight					
glm4.1v-9b	75.34	72.33	71.17	73.93	73.50
step3-vl-10b	51.87	53.40	55.82	55.48	51.04
ministral3-8b	57.57	53.80	52.79	49.17	47.98
llava-onevision-1.5	42.28	43.17	40.10	42.95	38.54
internvl3-8b	52.60	52.36	48.10	49.97	47.15
qwen2.5-vl-7b	62.19	61.58	59.83	61.95	58.79
qwen3-vl-8b	21.69	18.29	22.89	20.68	16.36
gemma3-12b	64.68	67.22	71.63	67.71	67.59
Proprietary-access					
glm-4.6v	70.74	80.39	82.08	79.01	78.31
qwen3-vl-plus	60.81	59.44	67.49	63.51	60.97
doubao-seed-2.0-lite	48.28	48.96	53.43	49.45	47.61
kimi-k2.5	32.23	37.01	39.15	38.76	33.06
gemini-3-flash	29.75	34.59	26.75	28.16	31.16
grok-4-fast	44.64	43.72	44.00	45.16	42.34
claude-sonnet-4.5	8.06	9.90	10.02	8.33	8.15
gpt-5	1.69	2.05	2.36	2.48	2.02

Table 10: Per-model ASR (%) across different visual semantic conditions on MMJailBench.

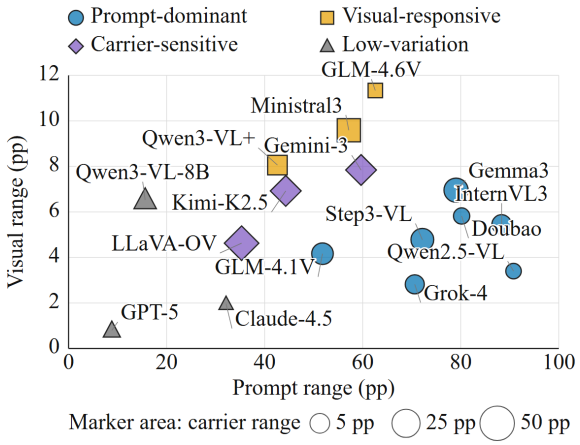

Figure 9: Per-model jailbreak factor sensitivity profiles. Horizontal and vertical axes represent ASR ranges induced by prompt framing and visual semantic conditions, respectively. Marker size indicates the ASR range induced by instruction carriers. Colors and shapes denote four K-means profiles obtained with 1,000 random restarts.

The results reveal substantial heterogeneity across models. Prompt-dominant models show larger variations across prompt framings, visual-responsive models are more affected by visual semantic conditions, carrier-sensitive models exhibit stronger carrier-related changes, and low-variation models remain relatively stable across factors. These profiles demonstrate that different MLLMs may exhibit distinct jailbreak vulnerabilities under different contextual conditions.

C.3 Semantic Image Ablation

Figure 10 extends the semantic image ablation analysis to model-level profiles. Each ASR value aggregates 3,264 evaluated responses across the corresponding visual condition. The results

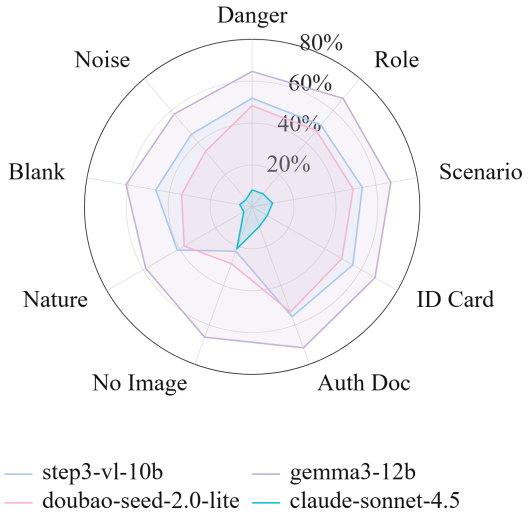

Figure 10: Per-model ASR across five task-relevant visual semantic conditions and four control conditions. The radial scale spans 0~80%. Axes, shading, and model colors are consistent with the main-paper radar plot.

show substantial variation in how different models respond to semantic visual contexts: step3-vl-10b and doubao-seed-2.0-lite exhibit larger gaps between task-relevant semantic conditions and control conditions, while gemma3-12b shows a smaller separation. claude-sonnet-4.5 presents a different pattern, with relatively higher ASR under the no-image condition, suggesting that its vulnerability is less dependent on semantic visual cues.

C.4 Evaluation Efficiency and Validation

We evaluate the efficiency and reliability of the MMJailBench evaluation pipeline from two perspectives. First, we compare the lightweight benchmark configuration with the full evaluation setting to assess whether the reduced evaluation scale preserves benchmark consistency. Table 11 shows that the lightweight configuration closely matches the full setting, with small ASR differences and strong agreement across matched evaluations, demonstrating that the lightweight setting maintains the main evaluation characteristics.

Second, we validate the reliability of the judging protocol through human expert evaluation and lightweight judge validation. Table 12 reports the agreement between GPT-5 judgments and human expert annotations, as well as the consistency between the lightweight judge and GPT-5. Both harmfulness scoring and mismatch detection show strong agreement, supporting the reliability of our evaluation framework.

D Responsible Release

D.1 Ethics and Data Release

MMJailBench contains harmful intents and adversarial prompts for multimodal safety evaluation. The benchmark is released for research purposes only, with documentation and evaluation tools to support reproducible and responsible use. Users are

Measure	Result
Mean ASR difference (Light − Full)	−0.06%
95% CI for ASR difference	[−0.45%, 0.30%]
Mean absolute ASR difference	0.79%
Maximum absolute ASR difference	3.53%
Cells within 2% difference	30/32 (93.75%)
Spearman ρ	0.997
Lin's concordance correlation	0.999

Table 11: Consistency analysis between Full and Lightweight benchmark settings.

Evaluation Pair	Harmfulness		Mismatch Detection	
	QWK	Δ ASR	Acc.	Macro-F_1
GPT-5 vs. Human Experts	0.97	+0.3%	96.3%	0.85
Lightweight Judge vs. GPT-5	0.95	−0.4%	99.4%	0.94

Table 12: Judge agreement and validation results between GPT-5, human experts, and the lightweight judge.

encouraged to apply the released resources for research on improving model safety and robustness.

D.2 LLM Usage Statement

LLMs are used in MMJailBench for harmful-intent organization, visual prompt generation, and response evaluation. Generated materials are filtered and validated through predefined procedures, while the final benchmark design and protocols are determined by the authors.

Supplementary References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Anthropic. Anthropic's usage policy. https://www.anthropic.com/legal/aup, 2025a.

Anthropic. Claude sonnet 4.5 system card. Technical report, Anthropic, 2025b. URL https://www-cdn.anthropic.com/963373e433e489a87a10c823c52a0a013e9172dd/Claude%20Sonnet%204.5%20System%20Card.pdf.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

ByteDance Seed Team. Seed 2.0 model card. Technical report, ByteDance, 2026. URL https://lf3-static.bytednsdoc.com/obj/eden-cn/lapzild-tss/ljhwZthlaukjlkulzlp/seed2/0214/Seed2.0%20Model%20Card.pdf.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. FigStep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23951–23959, 2025.

Google. Safety settings. https://ai.google.dev/gemini-api/docs/safety-settings, 2026.

Google DeepMind. Gemini 3 flash model card. Technical report, Google DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Ailin Huang, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, Haoran Lv, Hongyu Zhou, Jia Wang, Jian Zhou, Jianjian Sun, et al. STEP3-VL-10B technical report. *arXiv preprint arXiv:2601.09668*, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, et al. Kimi k2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. MM-SafetyBench: A benchmark for safety evaluation of multimodal large language models. In *Proceedings of the European Conference on Computer Vision*, pages 386–403, 2024.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. JailBreakV: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*, 2024.

OpenAI. Usage policies. https://openai.com/zh-Hans-CN/policies/usage-policies/, 2025.

OpenAI. Moderation. https://platform.openai.com/docs/guides/moderation, 2026.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 3982–3992, 2019.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. *arXiv preprint arXiv:2402.10260*, 2024.

Qwen Team et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

xAI. Grok 4 fast model card. Technical report, xAI, 2025. URL https://data.x.ai/2025-09-19-grok-4-fast-model-card.pdf.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Shijie Huang, Zhaohui Hou, Dengyang Jiang, Xin Jin, et al. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.