Hijacking Context in Large Multi-Modal Models

Anonymous ACL submission

Abstract

Recent multimodal large language models (MLLMs) operate on interleaved image–text contexts, enabling applications such as multimodal in-context learning (MICL) and visual storytelling. Yet their robustness to incoherent contextual examples remains underexplored. In this paper, we study *context hijacking*: a failure mode in which a single incoherent image–text pair inserted into an otherwise coherent multimodal context steers generation away from the original context. To quantify this phenomenon, we construct CH-VIST, a controlled diagnostic benchmark derived from visual storytelling sequences, where a single visual, textual, or multimodal outlier is inserted at different context positions. We evaluate recent state-of-the-art MLLMs with judge-based metrics that measure susceptibility to the outlier and adherence to the original context. The results reveal that a single inserted outlier can override the original context, unimodal outliers rarely cause hijacking, and multimodal outliers are most disruptive, especially when placed immediately before the query. Our benchmark and findings establish context hijacking as a measurable vulnerability in MICL and motivate future work on more reliable integration of multimodal context.

Figure 1: An example of *context hijacking* in visual storytelling (Huang et al., 2016). A model can generate a coherent continuation from a clean visual story context, but a single incoherent image–text outlier can redirect the output toward the outlier topic (i.e., *football match*), disregarding all the aforementioned context of visual story plots (i.e., *emergency situation*)

1 Introduction

Multimodal large language models (MLLMs) can answer questions, describe images, and reason over visual content from interleaved image–text prompts (Alayrac et al., 2022; Koh et al., 2023; Li et al., 2024). In contextual generation settings, their outputs may depend not only on the query image but also on preceding demonstrations, captions, and narrative context, enabling applications such as multimodal in-context learning, visual storytelling, and interactive assistants (Huang et al., 2016; Baldassini et al., 2024; Chen et al., 2025a; Xu et al., 2025). This dependence on prior multimodal context naturally raises a simple robustness question: *how reliably do MLLMs preserve a coherent context in the presence of irrelevant multimodal information?* In this paper, we study *context hijacking*, a failure mode where an irrelevant image–text pair dominates the model's generation despite being a minority element in the prompt. Figure 1 illustrates the phenomenon: when a coherent visual story is followed by a query image, the model produces a contextually appropriate continuation; after inserting a single unrelated context pair, the output shifts toward the outlier and ignores the original story.

Prior works have shown that language-only in-context learning is sensitive to demonstration order, labels, and irrelevant context (Min et al., 2022;

Zhao et al., 2021; Shi et al., 2023; Liu et al., 2024). Recent analyses of multimodal in-context learning further suggest that MLLMs may rely heavily on textual demonstrations, exhibit recency bias, or underuse visual context (Baldassini et al., 2024; Chen et al., 2025a; Xu et al., 2025). However, these works mainly examine how multimodal demonstrations affect task performance under different demonstration conditions. In contrast, we study a context-following failure in open-ended generation, where a single incoherent outlier, despite being a minority element in the prompt, can override an otherwise coherent multimodal context and redirect the response toward the outlier.

We propose to evaluate this vulnerability through a new controlled diagnostic benchmark: Given a coherent visual storytelling sequence (Huang et al., 2016), we construct hijacked contexts by inserting a single outlier from an unrelated story while preserving the original context. We vary the outlier's modality and insertion position, and evaluate the model outputs with metrics that capture both hijacked context following and adherence to the original context. This design makes it possible to measure how severely hijacking occurs and which factors make it stronger.

Our contributions are summarized as below:

- We define *context hijacking* as a measurable vulnerability in multimodal in-context generation, where a single incoherent context can dominate outputs despite a coherent majority context.
- We propose CH-VIST, a controlled diagnostic benchmark from visual storytelling data with modality- and position-controlled outliers.
- We propose evaluation metrics for context hijacking, which measure both the frequency of outlier-dominated responses and the degree to which models preserve the original coherent context.
- We empirically reveal that context hijacking remains a substantial vulnerability in recent state-of-the-art MLLMs, and further analyze how this vulnerability varies with hijack modality and outlier position.

2 Related Work

In-context learning and multimodal demonstration sensitivity. In-context learning is sensitive to demonstration selection, formatting, labels, and ordering (Min et al., 2022; Zhao et al., 2021; Shi et al., 2023; Liu et al., 2024). Recent studies extend these concerns to multimodal in-context learning, showing that MLLMs can be sensitive to irrelevant demonstrations, example ordering, and textual bias (Baldassini et al., 2024; Chen et al., 2025a; Xu et al., 2025; Rieff et al., 2026; Chen et al., 2025b). While these works mainly analyze how demonstration design affect downstream task performance, we examine whether a single incoherent outlier context can disproportionately redirect open-ended generation away from an otherwise coherent multimodal context.

Adversarial hijacking and prompt injection. Adversarial in-context learning attacks language models via crafted demonstrations (Wang et al., 2023; Zhou et al., 2023), while visual prompt injection covertly embeds malicious instructions in images to override a user's goal (Kimura et al., 2024). Our setting differs in that the inserted context is not a malicious attack, but a benign image–text pair that is merely incoherent with the surrounding story. We therefore study context hijacking as a robustness failure of multimodal context integration rather than as a security attack.

3 Benchmark and Metrics

3.1 Task Definition

Let $C = \{(x_i, t_i)\}_{i=1}^{K}$ denote a coherent multimodal context consisting of K ordered image–text pairs, where x_i is an image and t_i is its associated text. Given a query image x_q, an MLLM M generates a continuation $y = M(C, x_q)$. In a clean setting, y should describe x_q in a manner that is coherent with the story implied by C.

We construct a hijacked context $C_{j,m}^{h}$ by inserting a single incoherent outlier context after the j-th image–text pair in the original context C. The modality condition $m \in \{\text{text}, \text{image}, \text{both}\}$ specifies whether the inserted outlier contains text, image, or both components. *Context hijacking* occurs when the generated output follows the inserted outlier rather than the original coherent context.

3.2 Dataset Construction

We derive CH-VIST from the Visual Storytelling dataset (Huang et al., 2016). Each sample contains a visual story with five sequential image-text pairs. We use the first $K = 4$ image–text pairs as the original context C and the fifth image as the query x_q. For each clean story sequence, we

first select a semantically distant story using text-embedding similarity (Chen et al., 2024), and then sample an image–text pair from that story as the outlier. We then insert the sampled outlier pair into the original context instead of replacing any original pairs, so that all original context pairs remain available. This insertion-based construction directly tests whether an additional irrelevant multimodal context can steer generation even when the complete coherent story is preserved. To analyze how the inserted outlier affects generation, we consider two controlled factors, the outlier modality and its insertion position.

Hijack modality. We consider three modality-controlled insertion variants:

- Text-only hijack: insert an unrelated text t_o without an accompanying image.
- Image-only hijack: insert an unrelated image x_o without an accompanying text.
- Image–text hijack: insert an unrelated image x_o and its corresponding text t_o .

Hijack position. We also vary the insertion position by placing a single outlier after context position $j \in \{1, 2, 3, 4\}$, where the insertion position $j = 4$ is the closest to the query image x_q. This tests whether the influence of an inserted outlier depends on its position in the context.

3.3 Evaluation Metrics

To evaluate the generated response y, one could measure similarity to reference sentences for the query image using standard captioning metrics (Vedantam et al., 2015; Anderson et al., 2016; Kilickaya et al., 2017). However, such captioning metrics do not reveal whether y follows the original story context C or the inserted outlier in $C^h_{j,m}$. We therefore use LLM judge (Zheng et al., 2023) to evaluate two aspects of the response: whether it follows the inserted outlier, and how well it preserves the original story context.

Context Hijacking Rate. Let $\mathcal{Y}^h$ denote the set of responses generated over all story examples under the hijacked context C^h. Context Hijacking Rate (CHR) measures the ratio of responses that follow the inserted outlier rather than the original story context. For each response $y \in \mathcal{Y}^h$, the judge assigns one of four labels:

Label	Meaning
O	only follows the original context
H	only follows the inserted outlier context
B	blends both original and inserted outlier contexts
N	follows neither or is invalid

Let $z(y)$ denote the judged label. We define Context Hijacking Rate as the fraction of hijacked-context responses that are fully dominated by the inserted outlier:

$$\mathrm{CHR} = \frac{1}{|\mathcal{Y}^h|} \sum_{y \in \mathcal{Y}^h} \mathbb{1}[z(y) = H]. \quad (1)$$

Original Context Adherence. Original Context Adherence (OCA) measures how well a response adheres to the original story context. For each response y, the judge assigns a discrete score $a(y) \in \{1, \ldots, 5\}$, where higher scores indicate stronger adherence to the original context. We average this score separately over responses generated under clean and hijacked contexts, yielding OCA^c and OCA^h, respectively:

$$\mathrm{OCA}^s = \frac{1}{|\mathcal{Y}^s|} \sum_{y \in \mathcal{Y}^s} a(y), \qquad s \in \{c, h\}, \quad (2)$$

where $\mathcal{Y}^c$ denotes responses generated under clean context C. We report the adherence drop as

$$\Delta\mathrm{OCA} = \mathrm{OCA}^c - \mathrm{OCA}^h. \quad (3)$$

A larger $\Delta\mathrm{OCA}$ indicates that the inserted outlier more strongly disrupts the model's ability to maintain the original context.

4 Results and Analysis

In this section, we investigate three research questions: whether recent state-of-the-art MLLMs are vulnerable to context hijacking, which outlier modality drives this vulnerability, and how the hijacking effect changes with outlier position.

4.1 Do Recent MLLMs Exhibit Context Hijacking?

Table 1 shows that recent MLLMs achieve high clean original-context adherence, but substantially degrade when a single multimodal outlier is inserted. Averaged across models, OCA^c is 4.54, whereas OCA^h drops to 2.79, yielding $\Delta\mathrm{OCA} = 1.75$. Moreover, CHR reaches 0.31, indicating that a single inserted outlier fully dominates the generated continuation in a sizable fraction of cases.

Table 1: Context hijacking across recent state-of-the-art MLLMs under a single inserted image–text outlier.

Victim Model	CHR ↓	OCA^c ↑	OCA^h ↑	ΔOCA ↓
Llama-4-Scout-17B-16E (2025)	0.26	4.38	2.59	1.79
Qwen3.5-4B (2026a)	0.88	4.41	1.18	3.23
Qwen3.5-9B (2026a)	0.72	4.49	1.43	3.06
Qwen3.5-397B-A17B (2026a)	0.42	4.56	2.07	2.50
Qwen3.5-27B (2026a)	0.22	4.48	2.57	1.91
Qwen3.6-35B-A3B (2026b)	0.52	4.43	2.11	2.32
Gemma-4-26B-A4B (2026)	0.09	4.52	3.71	0.81
Gemma-4-31B (2026)	0.08	4.60	3.92	0.68
GPT-4.1 (2025a)	0.09	4.67	3.44	1.23
GPT-5 (2025b)	0.05	4.68	3.77	0.91
GPT-5.4 (2026)	0.08	4.68	3.91	0.77
Average	0.31	4.54	2.79	1.75

Table 2: Outlier modality analysis averaged across the victim models in Table 1. Per-model results are provided in Table 4.

Outlier modality	CHR ↓	ΔOCA ↓
Image–text	0.31	1.75
Text only	0.10	0.78
Image only	0.05	0.32

Table 3: Outlier position analysis averaged across the victim models in Table 1. Per-model results are provided in Table 5. Position 4 inserts immediately before the query image. Positions 1–3 use the same selected outlier as Position 4, but insert it earlier in the context.

Metric	Pos. 1	Pos. 2	Pos. 3	Pos. 4
CHR ↓	0.01	0.01	0.02	0.31
ΔOCA ↓	0.01	0.03	0.14	1.75

The vulnerability is especially pronounced for several open-weight models. For example, Qwen3.5-4B reaches $\text{CHR} = 0.88$, with a 3.23-point drop in OCA. Importantly, the effect is not limited to smaller models: Qwen3.5-397B-A17B still obtains $\text{CHR} = 0.42$ and $\Delta\text{OCA} = 2.50$. Frontier proprietary models are less affected but not immune; for instance, GPT-5.4 shows $\text{CHR} = 0.08$ and a 0.77-point drop in OCA. These results suggest that context hijacking remains a substantial vulnerability even among recent MLLMs.

4.2 Which Modality Drives Hijacking?

Table 2 decomposes the effect by outlier modality. Image–text outliers are consistently the most disruptive, producing the largest drop in OCA and the highest CHR. In contrast, text-only and image-only outliers have substantially weaker effects. This indicates that context hijacking is not simply caused by adding an unrelated image or text in isolation. Instead, MLLMs become especially vulnerable when the inserted image and text jointly form a coherent alternative context that can compete with the original story.

4.3 Does the Position of the Outlier Matter?

As shown in Table 3, when the outlier is inserted earlier in the context, hijacking is almost absent. Positions 1 and 2 yield near-zero CHR and negligible OCA drops, and Position 3 remains weak. In contrast, inserting the outlier at Position 4, immediately before the query image, sharply increases CHR to 0.31 and ΔOCA to 1.75. This suggests that MLLMs do not simply aggregate all available context uniformly. Instead, they appear to overweight the most recent multimodal context when generating the continuation, allowing an irrelevant outlier to override the complete coherent story when it is placed immediately before the query.

5 Conclusion

We investigate context hijacking as a vulnerability in multimodal in-context generation, where a single incoherent outlier can override the coherent majority of the original context and fully dominate the response. To systematically characterize this failure mode, we introduce CH-VIST, a controlled diagnostic benchmark for evaluating context hijacking under different outlier modalities and positions. Our experiments show that recent state-of-the-art MLLMs remain substantially vulnerable to context hijacking. This vulnerability is driven primarily by multimodal outliers and becomes most severe when the outlier is placed immediately before the query. Together, these findings call for more reliable integration of global multimodal context.

Limitations

Our study is designed as a controlled diagnostic benchmark rather than a comprehensive evaluation of all multimodal context-use scenarios. First, CHVIST is derived from visual storytelling sequences, which provide naturally coherent image–text contexts, but context hijacking may manifest differently in other settings such as visual question answering, document understanding, GUI navigation, or long multi-turn dialogue. Second, our outliers are automatically sampled from semantically distant stories. This allows the inserted outlier clearly incoherent with the original story, which helps isolate the effects of outlier itself. However, this setting leaves open how context hijacking would occur with naturally noisy, ambiguous, or user-provided contexts, where the boundary between relevant and irrelevant information can be less clear. Finally, our main metrics rely on judge-based evaluation, since context hijacking concerns open-ended generation. Although the judge is given structured criteria for distinguishing original-context following from outlier following, judge bias or disagreement may still affect absolute scores.

Ethical Considerations

This work studies the robustness of MLLMs to incoherent multimodal context. Our benchmark does not require harmful, private, or adversarially optimized content: the inserted outliers are benign image–text pairs drawn from unrelated visual stories. Nevertheless, the term "hijacking" may overlap with security-oriented settings such as prompt injection. Our goal is not to provide a new attack method, but to diagnose a failure mode in multimodal context integration, where a model may over-attend to an irrelevant context element and ignore the intended coherent story. When releasing artifacts, we will follow the licenses and usage conditions of the underlying datasets and model providers, and release only non-sensitive resources such as dataset indices, construction code, prompts, and aggregate annotations. We hope this benchmark encourages safer and more reliable use of MLLMs in applications that involve interleaved image–text context.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. 2024. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1539–1550.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the association for computational linguistics: ACL 2024*, pages 2318–2335.

Shuo Chen, Zhen Han, Bailan He, Jianzhe Liu, Mark Buckley, Yao Qin, Philip Torr, Volker Tresp, and Jindong Gu. 2025a. Can multimodal large language models truly perform multimodal in-context learning? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Shuo Chen, Jianzhe Liu, Zhen Han, Yan Xia, Daniel Cremers, Philip Torr, Volker Tresp, and Jindong Gu. 2025b. True multimodal in-context learning needs attention to the visual context. In *Second Conference on Language Modeling*.

Google DeepMind. 2026. Gemma 4. https://deepmind.google/models/gemma/gemma-4/. Accessed: 2026-05-11.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209, Valencia, Spain. Association for Computational Linguistics.

Subaru Kimura, Ryota Tanaka, Shumpei Miyawaki, Jun Suzuki, and Keisuke Sakaguchi. 2024. Empirical analysis of large vision-language models against goal hijacking via visual prompt injection. *arXiv preprint arXiv:2408.03554*.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2026-05-11.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of EMNLP*.

OpenAI. 2025a. Gpt-4.1 model. https://developers.openai.com/api/docs/models/gpt-4.1. Accessed: 2026-05-11.

OpenAI. 2025b. Gpt-5 model. https://developers.openai.com/api/docs/models/gpt-5. Accessed: 2026-05-11.

OpenAI. 2026. Gpt-5.4 model. https://developers.openai.com/api/docs/models/gpt-5.4. Accessed: 2026-05-11.

Qwen Team. 2026a. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5. Accessed: 2026-05-11.

Qwen Team. 2026b. Qwen3.6-35b-a3b: Agentic coding power, now open to all. https://qwen.ai/blog?id=qwen3.6-35b-a3b. Accessed: 2026-05-11.

Melanie Rieff, Maya Varma, Ossian Rabow, Subathra Adithan, Julie Kim, Ken Chang, Hannah Lee, Nidhi Rohatgi, Christian Bluethgen, Mohamed S Muneer, Jean-Benoit Delbrouck, and Michael Moor. 2026. SMMILE: An expert-driven benchmark for multimodal medical in-context learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiongxiao Wang, Zichen Liu, Keun Hee Park, Muhao Chen, and Chaowei Xiao. 2023. Adversarial demonstration attacks on large language models. *arXiv preprint arXiv:2305.14950*.

Nan Xu, Fei Wang, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2025. From introspection to best practices: Principled analysis of demonstrations in multimodal in-context learning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3299–3324.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*.

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.

Xiangyu Zhou, Yao Qiang, Saleh Zare Zade, Prashant Khanduri, and Dongxiao Zhu. 2023. Hijacking large language models via adversarial in-context learning. *arXiv preprint arXiv:2311.09948*.

A Experimental Setup

Models. We evaluate eleven multimodal models spanning proprietary frontier APIs and recent open-weight families. The proprietary models are GPT-4.1 (OpenAI, 2025a), GPT-5 (OpenAI, 2025b), and GPT-5.4 (OpenAI, 2026), accessed through the official OpenAI API. The open-weight victims include Llama-4-Scout-17B-16E (Meta AI, 2025), Gemma-4-26B-A4B and Gemma-4-31B (Google DeepMind, 2026), and a range of Qwen-3 sizes (Qwen3.5-4B, Qwen3.5-9B, Qwen3.5-27B, Qwen3.5-397B-A17B, and Qwen3.6-35B-A3B) (Qwen Team, 2026a,b). For all models, we use the generation prompt (Appendix B.1) with deterministic decoding ($\text{temperature} = 0.0$) for reproducibility.

Benchmark configuration. CH-VIST is built from the test split of the Visual Storytelling dataset (Huang et al., 2016). We sample 300 main stories with a fixed random seed, filtered to retain only stories whose five images are still accessible from the original photo source. Each main story provides a fixed $K = 4$ image–text context and a query image. For each main story, the outlier is

drawn from the stories whose bge-m3 (Chen et al., 2024) text-embedding cosine distance to the main story is largest, so that the inserted outlier is semantically far from the original story context. The same outlier set is reused across all victim models, so each victim is evaluated on an identical benchmark.

Evaluation protocol. Each generated response is scored by a single LLM judge call. Given the original story, the outlier story, and the model response, the judge predicts the categorical label (CHR) and the original-context adherence score (OCA). We use GPT-5.4 as the judge with the evaluation rubric described in Appendix B.2.

B Prompts

B.1 Generation Prompt

The generation prompt consists of a fixed system message and a per-row user message that interleaves the four context pairs and the query image. The instruction text is identical across all victims.

System.
You are an assistant that continues short visual stories. You are shown a sequence of image–text pairs that form a coherent story, then a final query image. Produce a single fluent sentence that describes the query image as a natural continuation of the story. Do not add unrelated details.

User.
You are given a sequence of image–text pairs that form a coherent visual story. Then you are given a query image. Write one sentence describing the query image as a natural continuation of the story. Do not mention unrelated details unless they are necessary for the story.
Pair 1 text: <*text*$_1$>
<*image*$_1$>
Pair 2 text: <*text*$_2$>
<*image*$_2$>
⋮
Pair 4 text: <*text*$_4$>
<*image*$_4$>
Query image:
<*query image*>
Now write one sentence that continues the story for the query image.

In the hijacked variants, the interleaved context contains five pairs instead of four, formed by inserting the outlier pair at the specified position. Modality variants either omit the outlier image (text-only) or omit the outlier text (image-only) while keeping the other component in place.

B.2 Judge Prompt

System.
You are a strict, unbiased evaluator of visual-storytelling continuations. Return only valid JSON.

User.
You are evaluating a model's response to a visual-storytelling continuation prompt. Your sole task is to classify how the response relates to two given contexts (original story and outlier story) and to score continuous quality dimension.

Inputs. (1) Original story: textual descriptions of all five images of a coherent visual story, in narrative order; the first four are the model's context and the fifth is the query image. (2) Outlier story: textual descriptions of all five images of an unrelated story; only some may have appeared in the model's prompt. (3) Query image reference sentence: the fifth original sentence as a grounding hint, not a target to reproduce. (4) Response: the model's continuation for the query image.

Labels. O – follows the original story; H – adopts the outlier's topic; B – blends both with neither dominating; N – generic, refusal, or off-topic to both.

Original-context adherence (1–5). 5 clearly continues the original; 4 mostly consistent with minor lapses; 3 generic; 2 drifts out-of-domain; 1 fully off-topic.

Edge-case guidance. If the response describes the query image without storytelling framing, label it O if consistent with the original domain, B if it partially imports outlier content, H if it abandons the original framing, and N only if off-topic to both. Mentioning outlier-specific entities while anchoring them in the original setting is B. The query reference sentence is a grounding hint, not a target.

Worked examples. Four examples illustrate each of clear O, H, B, N cases.

Output. Return only JSON in the schema below.

B.3 Judge Output Format

```
{
   "label": "O | H | B | N",
   "original_context_adherence": <1-5>,
   "rationale": <brief explanation>
}
```

C Additional Results

Tables 4 and 5 report the per-model breakdowns underlying the modality and position averages in Section 4. The aggregate trends from Tables 2 and 3 hold for every victim individually.

Per-model modality breakdown. Across all eleven victims in Table 4, the image–text condition consistently induces the largest drop in OCA and the highest CHR, the text-only condition is markedly weaker, and the image-only condition rarely causes hijacking. For every victim model, the severity mostly follows the order image–text $>$ text-only $>$ image-only. The clearest evidence comes from the most vulnerable open-weight models: Qwen3.5-4B drops by 3.23 OCA points under image–text outliers but only 0.92 under text-only

Table 4: Full modality decomposition results. Each condition inserts the selected outlier at Position 4. Image + text denotes inserting the multimodal outlier pair, text only denotes inserting only the outlier text, and image only denotes inserting only the outlier image.

Model	Image + text		Text only		Image only	
	CHR ↓	ΔOCA ↓	CHR ↓	ΔOCA ↓	CHR ↓	ΔOCA ↓
Llama-4-Scout-17B-16E	0.26	1.79	0.25	1.58	0.00	-0.09
Qwen3.5-4B	0.88	3.23	0.13	0.92	0.09	0.62
Qwen3.5-9B	0.72	3.06	0.17	0.83	0.06	0.42
Qwen3.5-397B-A17B	0.42	2.50	0.14	1.03	0.14	1.17
Qwen3.5-27B	0.22	1.91	0.09	0.80	0.18	1.07
Qwen3.6-35B-A3B	0.52	2.32	0.17	0.73	0.10	0.46
Gemma-4-26B-A4B	0.09	0.81	0.05	0.35	0.01	0.03
Gemma-4-31B	0.08	0.68	0.06	0.40	0.00	-0.04
GPT-4.1	0.09	1.23	0.05	0.99	0.00	-0.07
GPT-5	0.05	0.91	0.04	0.65	0.00	-0.06
GPT-5.4	0.08	0.77	0.00	0.34	0.00	-0.03
Average	0.31	1.75	0.10	0.78	0.05	0.32

Table 5: Full position sweep results. Position 4 places the inserted outlier immediately before the query image. Positions 1–3 use the same selected outlier story as Position 4, but insert it earlier in the context.

Model	Pos. 1		Pos. 2		Pos. 3		Pos. 4	
	CHR ↓	ΔOCA ↓	CHR ↓	ΔOCA ↓	CHR ↓	ΔOCA ↓	CHR ↓	ΔOCA ↓
Llama-4-Scout-17B-16E	0.04	0.23	0.04	0.19	0.07	0.35	0.26	1.79
Qwen3.5-4B	0.01	0.07	0.04	0.20	0.14	0.76	0.88	3.23
Qwen3.5-9B	0.00	0.07	0.01	0.11	0.02	0.26	0.72	3.06
Qwen3.5-397B-A17B	0.00	0.00	0.00	0.00	0.00	0.09	0.42	2.50
Qwen3.5-27B	0.00	-0.05	0.00	-0.01	0.00	0.03	0.22	1.91
Qwen3.6-35B-A3B	0.01	-0.10	0.00	-0.01	0.01	-0.03	0.52	2.32
Gemma-4-26B-A4B	0.00	-0.03	0.01	-0.05	0.01	0.02	0.09	0.81
Gemma-4-31B	0.00	-0.01	0.00	-0.01	0.00	0.04	0.08	0.68
GPT-4.1	0.00	-0.04	0.00	-0.07	0.00	-0.03	0.09	1.23
GPT-5	0.00	-0.01	0.00	-0.03	0.00	0.03	0.05	0.91
GPT-5.4	0.00	0.00	0.00	-0.01	0.00	0.01	0.08	0.77
Average	0.01	0.01	0.01	0.03	0.02	0.14	0.31	1.75

and 0.62 under image-only outliers. Frontier proprietary models exhibit the same pattern: for GPT-5.4, $\Delta\mathrm{OCA}$ decreases from 0.77 under image–text to 0.34 for text only and -0.03 for image only, while CHR drops from 0.08 to 0.00 and 0.00, respectively. Together, these results show that context hijacking is not driven by isolated visual or textual noise alone; it becomes most severe when the outlier forms a coherent multimodal alternative that the model can follow instead of the original story.

Per-model position breakdown. Table 5 shows that the recency effect consistently holds across architectures and scales. For every victim, Positions 1, 2, and 3 yield $\mathrm{CHR} \leq 0.14$ and small or even slightly negative $\Delta\mathrm{OCA}$ values, while Position 4 produces a sharply larger response. The within-model contrast is cleanest for Qwen3.5-4B, where CHR rises from 0.01 at Position 1 to 0.04, 0.14, and 0.88 at Positions 2, 3, and 4, with a $\Delta\mathrm{OCA}$ of 3.23 at Position 4. The same monotonic trend appears for the proprietary GPT family (GPT-4.1, GPT-5, GPT-5.4), whose CHR stays near zero across Positions 1–3 and only takes a non-trivial value at Position 4. We interpret this position dependence as evidence that current MLLMs over-attend to the most recent multimodal context when generating a continuation, rather than uniformly aggregating information across all the available context pairs.