Multimodal Task Interference: A Benchmark and Analysis of History-Target Mismatch in Multimodal LLMs

Masayuki Kawarada, Tatsuya Ishigaki, Hiroya Takamura
Artificial Intelligence Research Center, AIST
{kawarada.masayuki, ishigaki.tatsuya, takamura.hiroya}@aist.go.jp

Abstract

Task interference, the performance degradation caused by task switches within a single conversation, has been studied exclusively in text-only settings despite the growing prevalence of multimodal dialogue systems. We introduce a benchmark for evaluating this phenomenon in multimodal LLMs, covering six tasks across text and vision with systematic variation of history-target along three axes: modality mismatch, reasoning mismatch, and answer format mismatch. Experiments on both open-weights and proprietary models reveal that task interference is highly directional: switching from text-only to image-based targets causes severe performance drops, while the reverse transition yields minimal degradation. Interference is further amplified when mismatches co-occur across multiple dimensions, and is driven most strongly by modality differences, followed by answer format, while reasoning requirement shifts cause minimal degradation.

Keywords: large language models, multimodal, task interference, task switching, analysis

1. Introduction

Large language models (LLMs) have achieved strong performance across a wide range of tasks, including question answering, image captioning, and sentiment classification (Brown et al., 2020; OpenAI, 2023). In real-world dialogue scenarios, users frequently perform multiple tasks in succession within a single conversation. When the input transitions from one task (e.g., image captioning) to another (e.g., textual question answering), a phenomenon known as *task switching* occurs (see Figure 1 for an illustrative example). Such transitions can lead to a decline in model performance, an effect termed *task interference* (Gupta et al., 2024).

Although task interference has been documented in text-only settings, existing studies do not account for the multimodal nature of modern dialogue systems, which increasingly involve both text and images. This gap motivates the need for a dedicated evaluation framework that can systematically quantify how different types of mismatches between dialogue history and target inputs affect model performance. Beyond modality considerations alone, we hypothesize that the compatibility between reasoning requirements and answer format (e.g., classification versus generation) also plays a significant role in determining the degree of interference.

To address this need, we introduce a benchmark designed to evaluate task interference in multimodal LLMs. Our benchmark is built around three central research questions. First, how does a modality mismatch between history and target (e.g., transitioning from image-based to text-based input) affect model performance? Second, how does a reasoning mismatch, such as switching from commonsense question answering to factual recall, influence accuracy? Third, how does an answer format mismatch, such as transitioning from classification to generation, affect output quality?

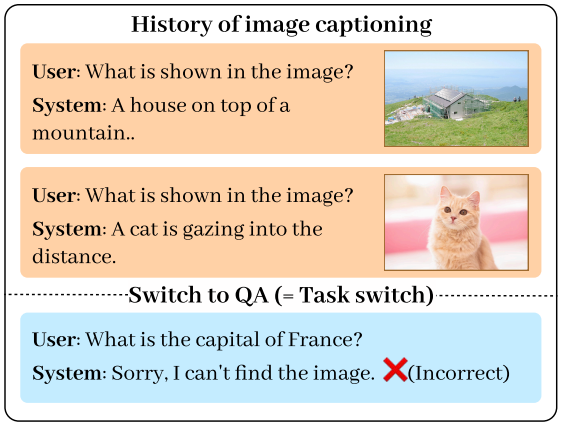

Figure 1: An illustrative example of multimodal task interference. After processing a conversational history of image captioning tasks, the model is prompted with a text-only question. The sudden task switch causes the model to erroneously expect a visual input, leading to a failure in answering a simple factual question.

The benchmark encompasses six diverse datasets covering sentiment classification, multiple-choice question answering, open-ended question answering, image captioning, and visual question answering. By constructing varied history-target configurations across these datasets, we enable controlled analysis of each mismatch dimension. We apply this benchmark to both open-weights and proprietary multimodal LLMs, allowing a direct comparison between the two categories under a unified evaluation protocol.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9282–9290
11–16 May 2026.

Dataset	Task Type	Modality ($\mathcal{M}$)	Reasoning ($\mathcal{R}$)	Answer Format ($\mathcal{A}$)
Rotten Tomatoes	Sentiment Classification	Text-only	No	Classification
MMLU	Multiple-choice QA	Text-only	Yes	Multiple-choice
TweetQA	Open-ended QA	Text-only	No	Generation
VQAv2	Visual QA	Image + Text	No	Short-answer
OK-VQA	Visual QA	Image + Text	Yes	Short-answer
COCO Captions	Image Captioning	Image + Text	No	Generation

Table 1: Overview of the six datasets used in our benchmark, categorized by the three axes of our task interference framework: Modality ($\mathcal{M}$), Reasoning requirement ($\mathcal{R}$), and Answer Format ($\mathcal{A}$).

Our results reveal that task interference in multimodal LLMs is highly directional and multifaceted. Cross-modal transitions exhibit a stark asymmetry: switching from text-only histories to image-based targets causes catastrophic performance drops, while the reverse yields minimal interference. Furthermore, simultaneous mismatches across multiple dimensions compound this degradation, demonstrating that modality mismatch alone cannot fully explain the interference. Models also show susceptibility to answer format changes but unexpected robustness to shifts in reasoning, highlighting the crucial role of structural and cognitive compatibility in dialogue stability.

Our contributions are: (1) a benchmark for evaluating task interference in multimodal LLMs along three axes (modality, reasoning, and answer format mismatch); (2) a comprehensive empirical study across six datasets with both open-weights and proprietary models; and (3) evidence that while modality switches cause significant interference, the performance degradation is most severe when compounded by simultaneous mismatches in reasoning requirements and answer formats.

2. Related Work

2.1. Multimodal Large Language Models

Multimodal large language models (MLLMs) extend LLMs to handle inputs across text and image modalities. Pioneering work such as Flamingo (Alayrac et al., 2022) demonstrated few-shot learning over interleaved image–text sequences, and subsequent systems including LLaVA (Liu et al., 2023), GPT-4 (OpenAI, 2023), and Qwen3-VL (Bai et al., 2025) have further advanced vision-language alignment and general-purpose assistant capabilities. As these models are increasingly deployed in multi-turn dialogue settings, the effect of accumulating heterogeneous conversational history becomes a practical concern. Prior work on long-context LLMs has shown that performance degrades non-uniformly as context grows, with information in the middle of long inputs being systematically overlooked (Liu et al., 2024). In multimodal dialogues, this challenge is further compounded by cross-modal history, motivating dedicated analysis of how different history compositions affect model behavior.

2.2. Task Interference

Task interference in LLMs has been studied both as a problem to mitigate and as a phenomenon to analyze. From the mitigation side, Chen et al. (2023) and Shen et al. (2024) propose mixture-of-LoRA architectures to reduce cross-task conflicts in MLLMs. From the analysis side, Gupta et al. (2024) formally define task interference as the performance degradation caused by task-switched conversational history in text-only settings, demonstrating significant accuracy drops across multiple task configurations. However, their analysis is limited to unimodal text settings and does not account for modality switches. Our work extends this line of research to multimodal dialogue by systematically evaluating interference along three axes of modality, reasoning requirement, and answer format, revealing interference patterns that text-only studies cannot capture, particularly a stark asymmetry in cross-modal transitions.

3. Task Interference

We formalize task interference in MLLMs. Let f be a model evaluated on a target task T_{tgt} with input x_{tgt} and reference y_{tgt}. In a conversational setting, the model is conditioned on a dialogue history $H = \{(x_i, y_i)\}_{i=1}^{N}$ of length N.

To systematically isolate the effect of *task switching*, we distinguish between two types of dialogue history:

- **Same-task History (H_{same})**: The history consists of examples from the target task itself ($T_H = T_{\text{tgt}}$), effectively acting as in-context learning.
- **Switched-task History (H_{switch})**: The history is sampled from a different task ($T_H \neq T_{\text{tgt}}$), introducing a task switch.

We quantify *Task Interference*, $\Delta^{(\%)}_{\text{switch}}$, as the relative performance degradation caused by a task switch compared to the ideal scenario where the context is consistent with the target task. Given an evaluation metric E, let E_{switch} and E_{same} be the performance scores under the switched-task and same-task histories, respectively:

$$E_{\text{switch}} = E(f(H_{\text{switch}}, x_{\text{tgt}}), y_{\text{tgt}}),$$
$$E_{\text{same}} = E(f(H_{\text{same}}, x_{\text{tgt}}), y_{\text{tgt}}),$$

where $f(H, x_{\text{tgt}})$ denotes the model output when conditioned on the dialogue history H followed by the target input x_{tgt}. We then define the switch cost as:

$$\Delta^{(\%)}_{\text{switch}} = 100 \cdot \frac{E_{\text{switch}} - E_{\text{same}}}{E_{\text{same}}}$$

Unlike prior studies (Gupta et al., 2024) that measure interference against a zero-shot baseline ($H = \emptyset$), our formulation isolates the specific impact of the *task switch*. By holding the presence of a conversation history constant, we control for confounding factors inherent to conversational prompting, such as increased context length, general in-context learning dynamics, and susceptibility to format failures. Task interference occurs when $\Delta^{(\%)}_{\text{switch}} < 0$, indicating that a switched-task history degrades performance compared to a relevant, same-task history.

To analyze the drivers of this interference, we characterize any task T as a tuple $T = \langle \mathcal{M}, \mathcal{R}, \mathcal{A} \rangle$, representing modality (e.g., text-only, image+text), reasoning requirement (e.g., factual recall, commonsense), and answer format (e.g., classification, generation). While a task switch occurs whenever the overall task changes ($T_H \neq T_{\text{tgt}}$), the specific attributes between the history and target tasks can independently match or differ. Therefore, within switched-task scenarios, we define the relationship along each axis as either a match (the attribute remains identical) or a mismatch (the attribute differs). Our benchmark evaluates interference by comparing these matched and mismatched conditions across three specific dimensions: modality ($\mathcal{M}_H = \mathcal{M}_{\text{tgt}}$ vs. $\mathcal{M}_H \neq \mathcal{M}_{\text{tgt}}$), reasoning ($\mathcal{R}_H = \mathcal{R}_{\text{tgt}}$ vs. $\mathcal{R}_H \neq \mathcal{R}_{\text{tgt}}$), and answer format ($\mathcal{A}_H = \mathcal{A}_{\text{tgt}}$ vs. $\mathcal{A}_H \neq \mathcal{A}_{\text{tgt}}$).

4. Experiments

We systematically evaluate model performance across all history-target, as detailed in the next section.

4.1. Target Tasks and Datasets

We use six benchmark datasets as shown in Table 1, spanning diverse task types and input modalities. Our dataset selection is guided by three criteria: (1) covering both text and image modalities to evaluate modality-specific and cross-modal interference, (2) controlling task difficulty by including tasks that require commonsense reasoning versus those that rely mainly on surface-level understanding, and (3) incorporating both classification/QA tasks with clear-cut answers and generation tasks where the output is inherently open-ended.

Table 1 shows the datasets classified in terms of these criteria. For text-based tasks, we use **Rotten Tomatoes** (Pang and Lee, 2005) for binary sentiment classification (commonsense not required), **Massive Multitask Language Understanding (MMLU)** (Hendrycks et al., 2021) for multiple-choice question answering in the mathematics domain (requiring multi-step reasoning), and **TweetQA** (Xiong et al., 2019) for open-ended QA over social media posts (commonsense not required). For image-based tasks, we include **OK-VQA** (Marino et al., 2019), which requires external commonsense knowledge beyond visual content, **VQAv2** (Goyal et al., 2017), which focuses on object recognition with minimal reasoning, and **COCO Captions** (Lin et al., 2014) for open-ended image caption generation. We examine LLM performance on all combinations of these datasets.

4.2. Multimodal Large Language Models

To comprehensively assess task interference, we evaluate four representative multimodal large language models, encompassing both proprietary API-based and open-weights architectures.

For the proprietary model, we evaluate **GPT-4.1-mini**[1] (OpenAI, 2023). For the open-weights models, we evaluate **Gemma-3n**[2] (Gemma Team, 2025), **Qwen3-VL**[3] (Yang et al., 2025), and **Pixtral**[4] (Agrawal et al., 2024). By including models with varying parameter scales and fusion mechanisms, we aim to determine whether susceptibility to task interference is a universal characteristic of multimodal systems or highly model-dependent.

4.3. Experimental Setup

For all evaluated models, the input prompts contain $N = 1, 3, 5$ randomly sampled history examples followed by a target input. For tasks involving images, visual inputs are passed through each model's native multimodal interface by substituting image-slot tokens in the prompt template with base64-encoded image representations.

[1] `gpt-4.1-mini-2025-04-14`
[2] `google/gemma-3n-E4B-it`
[3] `Qwen/Qwen3-VL-30B-A3B-Instruct`
[4] `mistralai/Pixtral-12B-2409`

	Modality (%)			Reasoning (%)			Answer Format (%)		
	mismatch	match	Δ	mismatch	match	Δ	mismatch	match	Δ
N=1									
GPT-4.1-mini	-8.89	-7.65	-1.24	-5.35	-11.88	6.53	-8.44	-8.12	-0.32
Gemma-3n	-10.49	-11.48	0.99	-10.84	-10.94	0.10	-11.11	-9.39	-1.72
Qwen3-VL	-9.50	2.65	-12.15***	-4.29	-5.03	0.74	-5.35	-0.03	-5.31**
Pixtral	-6.15	-1.34	-4.81***	-4.05	-4.42	0.38	-3.83	-6.75	2.91
N=3									
GPT-4.1-mini	-16.63	-11.55	-5.09**	-13.49	-15.87	2.38	-15.41	-9.33	-6.08**
Gemma-3n	-22.79	-20.50	-2.29*	-22.05	-21.67	-0.38	-21.94	-21.43	-0.52
Qwen3-VL	-14.58	-1.02	-13.56***	-7.74	-10.78	3.05	-8.96	-10.41	1.45
Pixtral	-8.54	-2.22	-6.31***	-5.17	-6.97	1.80	-5.66	-8.30	2.64
N=5									
GPT-4.1-mini	-19.99	-14.02	-5.97**	-16.25	-19.15	2.91	-18.48	-11.90	-6.58**
Gemma-3n	-25.25	-21.66	-3.60*	-23.68	-23.97	0.29	-24.25	-21.02	-3.23
Qwen3-VL	-14.87	-3.18	-11.69***	-8.93	-11.63	2.70	-10.45	-8.52	-1.93
Pixtral	-11.09	-4.39	-6.69***	-8.14	-8.72	0.58	-8.09	-10.46	2.37

Table 2: Axis-wise means (%) for mismatch and match groups, and their difference (Δ = mismatch - match). Significance marks on Δ: Welch's t-test ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

To investigate the precise impact of a task switch, we adopt a teacher-forcing approach when constructing the history prompts. Specifically, the response for each history example is taken directly from the ground-truth label or reference text. This methodology ensures that the dialogue context is perfectly accurate and eliminates the confounding factor of model generation errors propagating through the conversation.

To ensure the reproducibility and statistical robustness of our results, all experiments are conducted across five different random seeds for each condition. For each seed, we resample the history examples to account for variance in in-context example selection. Throughout these trials, the decoding temperature is set to 0 to eliminate stochasticity in model outputs, ensuring that observed variance reflects only the effect of history composition.

For the open-weights models, all local inference was conducted on a single NVIDIA H200 GPU using the vLLM inference framework (Kwon et al., 2023). For evaluation, we use the initial 1,000 instances from the test split of each target dataset.

4.4. Evaluation Metrics

To assess model performance, we select primary evaluation metrics tailored to the specific output format of each dataset. For classification and multiple-choice question answering tasks, specifically Rotten Tomatoes and MMLU, we report standard accuracy. For open-ended textual question answering on TweetQA, we evaluate the responses using the F1 score to measure the token-level overlap with the reference answers.

For visual question answering tasks, encompassing VQAv2 and OK-VQA, we employ the standard VQA accuracy metric to account for human annotator variance. Finally, for the open-ended image captioning task on COCO Captions, we use the CIDEr metric (Vedantam et al., 2015), which effectively evaluates the consensus between the generated caption and human reference captions.

Since the absolute scales of these metrics vary significantly (e.g., CIDEr scores versus standard percentages), we evaluate the switch cost using the relative percentage change ($\Delta^{(\%)}_{\text{switch}}$) as defined in Section 3. This normalization enables a unified and fair comparison of task interference across all datasets and diverse metric types.

5. Results and Discussion

We organize our findings according to three hypothesized factors contributing to task interference: (1) modality mismatch, (2) reasoning mismatch, and (3) answer format mismatch.

5.1. Effects of Task Interference along Three Axes

Table 2 presents the performance changes across the three mismatch axes. We observe a consistent trend where performance generally degrades even in the "match" conditions. This demonstrates that the mere occurrence of a task switch, shifting from one dataset to another while preserving the same modality, reasoning requirement, or answer format, inherently harms model accuracy. This baseline degradation strongly aligns with prior findings on task interference in text-only dialogue settings (Gupta et al., 2024). Building upon this observation, we dissect the additional interference induced

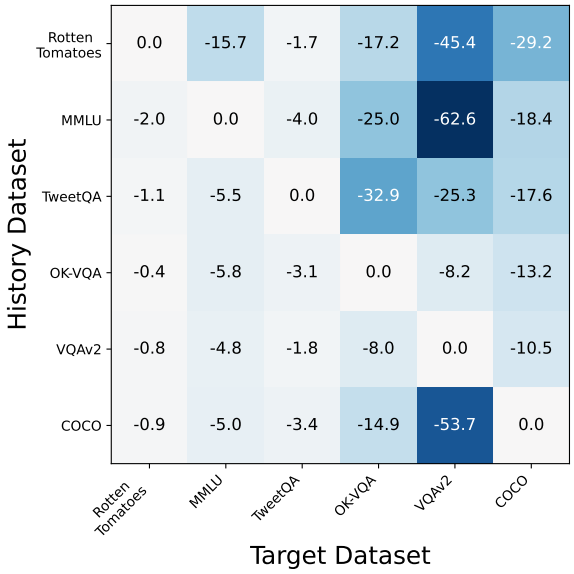

Figure 2: A heatmap visualizing the performance drop (relative change in %) across all pairwise combinations of history and target datasets for GPT-4.1-mini with a history length of $N = 3$.

specifically by mismatches along each of our three proposed axes.

Modality Mismatch Modality mismatch generally induces severe task interference, often being the most dominant factor across the evaluated models. While the impact varies in the single-shot setting ($N = 1$), increasing the history length to $N = 3$ and $N = 5$ reveals significant performance degradation for all models. For instance, Qwen3-VL experiences a massive drop with a Δ of -12.15% even at $N = 1$. Similarly, as the context lengthens to $N = 5$, GPT-4.1-mini and Gemma-3n show significant decreases of -5.97% and -3.60%, respectively. These results suggest that multimodal models struggle heavily to adapt when the input modality shifts between the dialogue history and the target prompt.

Reasoning Mismatch Interestingly, the data indicates that models are highly robust to shifts in reasoning requirements. Across almost all models and history lengths, the Δ values for reasoning mismatch are positive or near zero, and none reach statistical significance for negative degradation. For example, GPT-4.1-mini demonstrates a positive Δ of 6.53% at $N = 1$. This implies that switching between different cognitive tasks does not penalize model performance. In fact, maintaining the exact same reasoning type in the history might occasionally lead to over-conditioning on specific patterns, resulting in slightly worse outcomes than a mismatched history.

	Text→Image (%)	Image→Text (%)
N=1		
GPT-4.1-mini	-18.65***	0.88
Gemma-3n	-18.87***	-2.11***
Qwen3-VL	-19.84***	0.85
Pixtral	-12.94***	0.64
N=3		
GPT-4.1-mini	-30.38***	-2.88***
Gemma-3n	-39.91***	-5.67***
Qwen3-VL	-29.44***	0.28
Pixtral	-16.08***	-0.99
N=5		
GPT-4.1-mini	-35.48***	-4.50***
Gemma-3n	-42.70***	-7.81***
Qwen3-VL	-28.96***	-0.78***
Pixtral	-19.02***	-3.15***

Table 3: Directional modality transitions. Values are mean relative change (%) for each transition type. Significance markers are based on one-sample t-test against 0 for each transition ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Answer Format Mismatch The effect of answer format mismatch is evident but highly model-dependent. At $N = 1$, only Qwen3-VL shows a statistically significant drop of -5.31%. However, as the dialogue context grows, GPT-4.1-mini becomes notably susceptible to format changes, exhibiting significant performance drops of -6.08% at $N = 3$ and -6.58% at $N = 5$. Conversely, open-weights models like Pixtral and Gemma-3n remain relatively unaffected by shifts in the expected answer format across all context lengths. This indicates that while format interference exists, it does not universally degrade performance in the same destructive manner as modality switching.

5.2. Interaction of Multiple Mismatches

To better understand how the different dimensions of task interference interact, Figure 2 visualizes the performance drop across all pairwise history and target dataset combinations for GPT-4.1-mini at $N = 3$.

First, cross-modal switches show a strong asymmetry. Transitions from text-only history to image-based targets suffer severe degradation, peaking at a 62.6% drop from MMLU to VQAv2. Conversely, switching from image-based history to text-only targets yields marginal degradation, such as a mere 4.8% drop from VQAv2 to MMLU. Second, simultaneous mismatches across multiple dimensions appear to contribute to a compound interference effect. The massive drop from MMLU to VQAv2 is likely exacerbated by the fact that the modality, reasoning, and answer format all change at once. Furthermore, even within the same modality, chang-

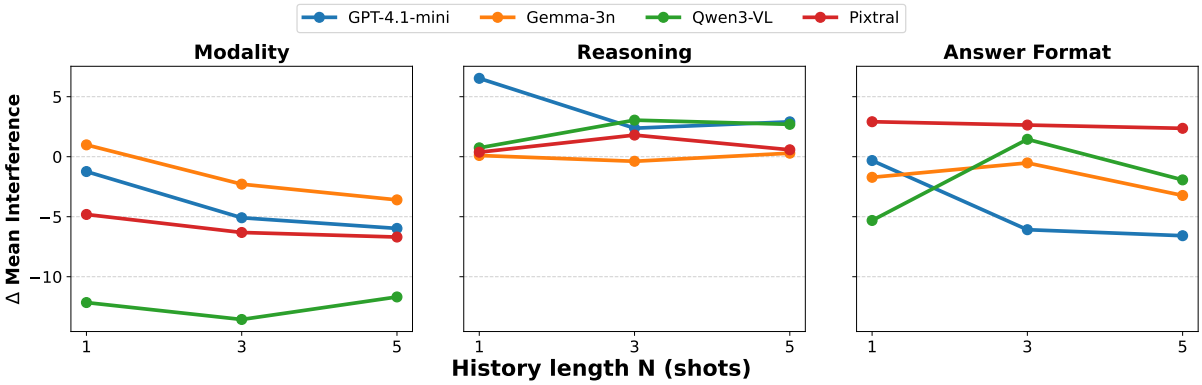

Figure 3: Performance difference (Δ) between mismatch and match conditions across varying history lengths ($N = 1, 3, 5$) for modality, reasoning, and answer format dimensions.

ing the reasoning and format requirements can be associated with significant interference, as seen in the 15.7% drop from Rotten Tomatoes to MMLU.

5.3. Asymmetry in Modality Transitions

While Section 5.1 established that modality mismatch significantly degrades overall performance, a closer examination reveals that this interference is highly directional. As shown in Table 3, there is a stark asymmetry between transitioning from text to image tasks and the reverse direction.

When models are conditioned on a text-only dialogue history and then prompted with a target image+text input, they suffer catastrophic performance drops. For example, as the context length increases to $N = 5$, the Text→Image transition results in massive decreases ranging from -19.02% in Pixtral to -42.70% in Gemma-3n. All evaluated models show highly statistically significant degradation in this specific direction.

Conversely, switching from an image-based history to a purely textual target task yields minimal interference. In the single-shot setting ($N = 1$), models like GPT-4.1-mini and Qwen3-VL exhibit slightly positive relative changes of 0.88% and 0.85%, respectively. Even at the longest context length ($N = 5$), the degradation in the Image→Text direction remains mostly in the single digits. This strong asymmetry suggests that accumulating a long textual context overwhelms the visual processing capabilities of the models. They appear to lock into a text-only reasoning mode, causing them to neglect or severely misinterpret the newly introduced visual input.

5.4. Impact of History Length

To understand how the accumulation of context affects task switching, Figure 3 illustrates the performance difference (Δ) across varying history lengths ($N = 1, 3, 5$) for each mismatch dimension. The trajectories reveal that the effect of history length is highly dependent on the type of mismatch.

In the case of modality mismatch, extending the history length generally exacerbates the interference. For models like GPT-4.1-mini, Gemma-3n, and Pixtral, the Δ becomes increasingly negative as the number of shots increases from $N = 1$ to $N = 5$. This steady downward trend indicates that a longer exposure to a specific modality strongly anchors the model's attention, making it progressively more difficult to process a sudden cross-modal target.

Conversely, the reasoning dimension exhibits remarkable stability across varying context lengths. The performance curves remain relatively flat and hover near or above the zero mark for all evaluated models. This visual evidence confirms that accumulating more examples of a mismatched reasoning type does not compound the cognitive interference, allowing the models to seamlessly adapt to the target prompt's reasoning requirement regardless of the context length.

Finally, the answer format dimension shows model-specific sensitivities to history length. While open-weights models maintain a relatively stable performance difference across varying N, GPT-4.1-mini displays a sharp performance drop as the context expands from $N = 1$ to $N = 3$. This suggests that certain models become easily locked into the answer format established by a longer dialogue history, severely hindering their ability to adapt to a different target format.

5.5. Qualitative Analysis of Task Interference Mechanisms

To qualitatively understand the interference mechanism, we manually analyzed the generated errors and observed a specific type of output-style drift induced by task-switched history. We observe clear output-style drift when visual target tasks are pre-

ceded by text-only history. For short-answer visual QA (VQAv2/OK-VQA), while same-task visual history leads to concise, task-appropriate answers, switched-task text-only history often yields verbose or reformulated outputs that deviate from the expected gold labels.

For example, in OK-VQA, the same-task output is a concise "soccer", but MMLU history switches the model's response to the more descriptive but non-matching "playing catch." Similarly, the same-task answer "ball" shifts to "frisbee" under Rotten Tomatoes history. Even when the semantics are largely preserved, formatting drift can lead to evaluation failures: in VQAv2, the same-task output is a simple "yes", while MMLU history induces the redundant sentence "Yes, the leaves are large."

Quantitatively, this bias is reflected in the average output length for VQAv2 targets, which increases from 1.69 words under same-task history to 3.69 words under MMLU history for GPT-4.1-mini. These results suggest that long textual contexts anchor the model into an elaborative completion mode, which is fundamentally incompatible with constrained multimodal evaluation metrics.

6. Conclusion

This paper investigated task interference in multimodal large language models. Our evaluation reveals that performance degradation is highly directional, exhibiting a stark asymmetry. Transitioning from text-only histories to image-based targets severely degrades performance, whereas the reverse causes minimal disruption. Furthermore, task interference compounds when models face simultaneous mismatches across multiple dimensions. The most catastrophic drops occur when modality, reasoning, and answer format change at once. Interestingly, models are vulnerable to modality and answer format shifts but remain unexpectedly robust to changes in reasoning requirements.

These findings demonstrate that modality matching alone cannot guarantee conversational stability. To build truly robust multimodal dialogue systems, future work must focus on dynamic context management, interference detection, and mixed-task instruction tuning.

Limitations

Our study has several limitations that highlight directions for future work. First, while we evaluate four representative MLLMs, our analysis does not cover the largest flagship models (e.g., full-scale GPT-4.1 or 70B+ parameter models). Investigating how model scale affects robustness remains an important next step.

Second, our experiments focus on short dialogue histories ($N \leq 5$) with a single task switch, whereas real-world conversations involve long contexts and multiple transitions. Third, the benchmark is currently restricted to text and image modalities, excluding emerging audio and video inputs.

Finally, while our teacher-forcing approach (using ground-truth history) successfully isolates the effects of task switching, it does not capture self-induced interference, which refers to the cascading effect whereby a model's own generation errors in prior turns corrupt subsequent context. In real-world scenarios, a model's previous generation errors or hallucinations can propagate, likely causing more severe performance degradation than reported here.

Acknowledgements

This paper is based on results obtained from AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain". We used ABCI 3.0 provided by AIST and AIST Solutions with support from "ABCI 3.0 Development Acceleration Use".

Bibliographical References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. 2024. Pixtral 12b.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and

Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. Qwen3-vl technical report.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Zeren Chen, Ziqin Wang, Zhen Wang, Huayang Liu, Zhenfei Yin, Si Liu, Lu Sheng, Wanli Ouyang, Yu Qiao, and Jing Shao. 2023. Octavius: Mitigating task interference in mllms via lora-moe. *arXiv preprint arXiv:2311.02684*.

Gemma Team. 2025. Gemma 3 technical report.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334. IEEE Computer Society.

Akash Gupta, Ivaxi Sheth, Vyas Raina, Mark Gales, and Mario Fritz. 2024. Llm task interference: An initial study on the impact of task-switch in conversational history. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14633–14652.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*. OpenReview.net.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014 – 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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*, 12:157–173.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204. Computer Vision Foundation / IEEE.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Ying Shen, Zhiyang Xu, Qifan Wang, Yu Cheng, Wenpeng Yin, and Lifu Huang. 2024. Multimodal instruction tuning with conditional mixture of lora. *arXiv preprint arXiv:2402.15896*.

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)*.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Tweetqa: A social media focused question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5020–5031, Florence, Italy. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.