COMPOSE AND FUSE: REVISITING THE FOUNDATIONAL BOTTLENECKS IN MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) promise enhanced reasoning by integrating diverse inputs such as text, vision, and audio. Yet, despite their perceptual strengths, their reasoning ability across modalities remains underexplored, with conflicting reports on whether additional modalities help or harm performance. These inconsistencies stem from a lack of controlled evaluation frameworks and analysis of models' internals to isolate *when* and *why* modality interactions support or undermine reasoning. We address this gap through a logic-grounded evaluation framework that categorizes multimodal reasoning into six interaction patterns, varying how facts are distributed across modalities and logically combined. Empirically, additional modalities enhance reasoning only when they provide independent and sufficient reasoning paths, while redundant or chained entailment support often hurts performance. Besides, models recognize cross-modal facts reliably and always reason on text effectively. Moreover, reasoning degrades in three systematic ways: weaker modalities drag down overall performance, conflicts bias preference toward certain modalities, and joint signals from different modalities fail to be integrated effectively. Therefore, we identify two core failures: *task-composition bottleneck*, where recognition and reasoning cannot be jointly executed in one pass, and *fusion bottleneck*, where early integration introduces bias. For further investigation, we find that attention patterns fail to encode fact usefulness, but a simple two-step prompting (recognize then reason) restores performance, confirming the task-composition bottleneck. Moreover, modality identity remains recoverable in early layers, and softening attention in early fusion improves reasoning, highlighting biased fusion as another failure mode. Overall, our findings show that integration, not perception, is the main barrier to multimodal reasoning, suggesting composition-aware training and early fusion control as promising directions.

1 INTRODUCTION

Multimodal large language models (MLLMs) extend traditional language models beyond text to incorporate additional modalities such as vision and audio (Li et al., 2025c; Xu et al., 2025; Yu et al., 2025; Abouelenin et al., 2025). By integrating complementary signals, MLLMs can form richer and more grounded representations of the world. Text offers structured and abstract information, audio encodes temporal and prosodic cues, and images convey spatial and visual context, together capturing facets of meaning that no single modality can express in isolation (Clark & Brennan, 1991; Mayer, 2002). Through such cross-modal integration, MLLMs aim to overcome the limitations of unimodal systems, enabling more robust understanding, stronger perceptual grounding, and support for more complex reasoning (Li et al., 2023; Bie et al., 2025; Raza et al., 2025; Coburn et al., 2025).

While MLLMs hold promise for enhanced reasoning by integrating diverse signals, the precise influence of additional modalities remains unclear, especially under complex reasoning scenarios. Existing studies offer conflicting observations: some report that incorporating vision or audio can improve model performance (Li et al., 2023; Guan et al., 2024; Fu et al., 2025), while others suggest that additional modalities introduce interference or confusion (Bie et al., 2025; He et al., 2025; Hou et al., 2025). However, these findings are often anecdotal or domain-specific, lacking a unified framework to systematically assess when and how multimodal input contributes to or undermines reasoning (Gupta et al., 2024; Coburn et al., 2025; Hao et al., 2025; Li et al., 2025b; Bi et al., 2025). In particular, it still remains unclear under what conditions additional signals strengthen reasoning,

add little, or actively impede it (Wu et al., 2025; Zhang et al., 2025). As a result, the role of modality interaction in reasoning, whether beneficial or detrimental, remains underexplored.

Moreover, most evaluations treat MLLMs as black-box systems, emphasizing external performance while leaving their internal mechanisms poorly understood (Liu et al., 2024b; Liang et al., 2023; Li et al., 2024). Even when empirical patterns emerge, such as degraded reasoning with added modalities, they are rarely accompanied by interpretability analyses that examine how models internally encode modality identity, assess evidence relevance, or perform cross-modal integration (Peng et al., 2025; Sinha et al., 2024; Yu et al., 2024; Wadekar et al., 2024). One contributing factor could lie in how these models are trained: current MLLMs are typically optimized using alignment-style objectives that pair vision or audio with text through paired supervision, contrastive learning, or instruction tuning (Zhao et al., 2024; Lin et al., 2024; Xie & Wu, 2024; Jiang et al., 2025). These objectives prioritize perceptual matching over cognitive composition, reinforcing shallow correlations rather than fostering deeper reasoning. As a result, while MLLMs often perform well on perception-heavy tasks (Li et al., 2023; Liu et al., 2024a), they struggle to generalize when reasoning demands flexible integration of multimodal information. Without interpreting the internal representations and fusion behaviors, it remains difficult to pinpoint where these limitations arise or how they might be overcome.

Prior evaluations report mixed effects of adding modalities to reasoning because they rarely control *where* decision-relevant facts appear or *how* those facts must be logically combined. We focus on *logical reasoning using information from multiple modalities*: how models use cross-modal facts to infer answers, rather than simple perception of unimodal content. To make effects measurable, we introduce six canonical interaction types (§ 2), grounded in propositional logic, that jointly vary (i) where the crucial facts are placed across modalities and (ii) how those facts must be combined to solve the task: *Equivalence* ($\equiv$, redundant encoding), *Alternative* ($\vee$, distinct but individually sufficient paths), *Entailment* ($\rightarrow$, chained support across modalities), *Independence* ($\emptyset$, a single modality carries the relevant fact), *Contradictory* ($\oplus$, mutually exclusive conclusions), and *Complementary* ($\wedge$, jointly necessary pieces). Instantiated through controlled, synthetic multiple choice reasoning tasks, this framework allows us to assess not only *when* added modalities that help or hurt, but also *why*.

To understand *when* added modalities help or hurt reasoning, we analyze performance across our six interactions. *Alternative* yields slight gains: added modalities help when they provide independent, individually sufficient reasoning paths. *Equivalence* offers no benefit in the presence of a strong modality (e.g., text), suggesting that redundant perceptual support rarely improves high unimodal performance. *Entailment* consistently degrades accuracy, showing that splitting multi-hop reasoning chains across modalities makes inference brittle (§ 3.2). *Independence* reveals performance bias, where reasoning accuracy heavily depends on which modality carries the decisive fact. *Contradictory* exposes preference bias, as models could favor certain modalities when inputs conflict. *Complementary* highlights a weakness in fusion, where models struggle to integrate necessary signals jointly (§ 3.2). Moreover, across all settings, text-only baselines approach ceiling and models reliably recognize facts across modalities. Therefore, we identify two core bottlenecks: (**i**) *task composition*: models struggle to jointly perform recognition and reasoning when information is split across modalities; and (**ii**) *multi-source fusion*: models lack robust mechanisms to select, weight, and combine heterogeneous information, leading to performance, preference, and fusion biases (§ 3.4).

To explain *why* these bottlenecks arise, we probe the internal behaviors of MLLMs in a controlled setting. First, although models recognize facts well and reason effectively (in text), their internal attention patterns fail to encode *usefulness* (i.e., distinguishing relevant facts from distractors). As a result, accuracy drops when recognition and reasoning must be composed within a single step, revealing a *task-composition bottleneck*. Explicitly decoupling the two stages through a two-step prompt substantially alleviates this issue (§ 4.1). Second, models preserve modality identity throughout processing, with the strongest signal concentrated in early decoder layers, but this preservation leads to biased weighting across modalities. Targeted interventions in these layers (e.g., softening early attention via increased temperature) significantly improve cross-modal reasoning, whereas modifications at later layers have little effect. This pattern confirms a *fusion bottleneck*, where biased early integration prevents balanced use of multimodal evidence (§ 4.2). Together, these insights indicate that additional modalities yield limited and often inconsistent benefits, with failures rooted not in perception but in integration. This calls for models that incorporate composition-aware training, supervision for evidence selection, and architectural mechanisms for early fusion control, so that extra modalities become assets for reasoning rather than sources of interference.

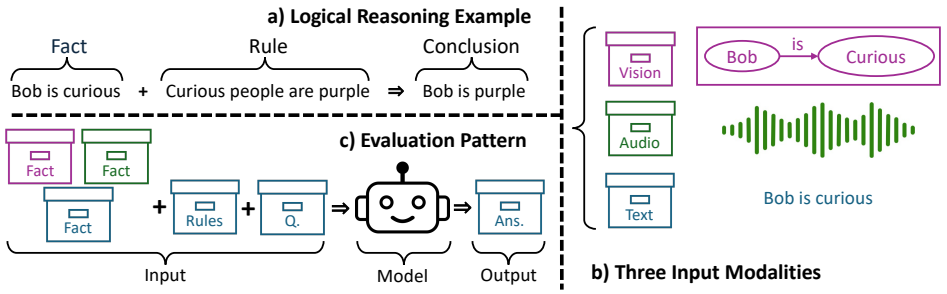


Figure 1: **Multimodal logical reasoning setup and evaluation pipeline. (a) Logical reasoning example:** a single-step deduction where the fact "*Bob is curious*" and the rule "*Curious people are purple*" entail the conclusion "*Bob is purple*." **(b) Modality renderings:** the same fact is rendered as text ("Bob is curious"), as audio via neural TTS, and as a schematic visual using graph visualization. **(c) Evaluation prompt pattern:** the model receives modality-specific fact blocks (text, audio, vision), followed by the rule set and the question with multiple-choice options; the model outputs the predicted answer.

2 Categorization: Modality Interaction in Reasoning

This section introduces our evaluation framework for multimodal logical reasoning. We first describe the general task setup: how facts, rules, and questions are constructed and how facts are rendered across modalities (§ 2.1). We then introduce details of the six canonical interaction types, which specify how useful information is distributed across modalities (§ 2.2). These settings allow us to systematically assess whether and how models integrate, ignore, or prioritize cross-modal evidence.

2.1 Multimodal Logical Reasoning Setup

Fig. 1 illustrates our reasoning task setup. Each instance consists of a set of facts, a set of rules (always in text), and a multiple-choice question. Facts convey information through different modalities: text, audio, or vision, and we control which facts are assigned to which modality (Fig. 1.b).

(a) Logical reasoning example. To isolate modality interaction, we adopt a simplified single-step reasoning setting inspired by Clark et al. (2020). Full details of the data construction are in App. B.1. For example, given the fact "*Bob is curious*" and the rule "*Curious people are purple*," the model should infer "*Bob is purple*." This setup avoids multi-hop complexity and directly tests the model's ability to identify and utilize relevant information across modalities.

(b) Three input modalities. To minimize confounds from low-level perception, we encode each fact in three controlled modalities (see Fig. 2b): (i) a short text sentence (e.g., "Bob is curious"), (ii) audio synthesized via neural TTS,[1] and (iii) a schematic visual rendered using graph visualization.[2] These controlled renderings ensure interpretability and reduce variability due to acoustic or visual complexity, allowing us to focus on reasoning and modality integration.

(c) Evaluation prompt pattern. Each prompt presents a randomized set of fact blocks from different modalities, followed by the text-based rule set and a multiple-choice question. To assess robustness and bias, we also inject *noisy facts* (irrelevant distractors) into the input. In the next subsection, we define six modality interaction types that vary how decision-relevant facts are distributed, enabling us to test different forms of cross-modal fusion, redundancy, and conflict.

2.2 Interaction Types

Equivalence (≡). All modalities redundantly encode the same fact (see Fig. 4 in App. B.3 for more details). For example, "*Erin is friendly*" appears in vision, audio, and text. This tests whether redundancy helps or harms reasoning, revealing models' ability to aggregate repeated evidence.

Alternative (∨). Each modality presents a different fact, but all facts independently satisfy a disjunctive rule (Fig. 5). For example, the rule "*Friendly person is clean. Purple person is clean. Red person is clean.*" is matched by "*Erin is friendly*", "*Erin is person*", and "*Erin is red*", in separate

[1] We use CosyVoice2 TTS to convert text to speech.

[2] We use GraphViz to generate simple entity-attribute diagrams.

modalities. This setting assesses whether models can leverage distinct but semantically aligned reasoning paths.

Entailment (→). Facts are distributed across modalities to form a multi-hop reasoning chain (Fig. 6). For example, "*Erin is bouncy*" → "*Erin is bright*" → "*Erin is friendly*" → "*Erin is purple*", with each step in a different modality. Only the final fact directly supports the answer. This setting probes models' capacity for chained inference across modalities.

Independence (∅). Only one modality contains the decision-relevant fact, while the others include unrelated distracting facts (Fig. 7). For example, only the vision modality presents "*Erin is friendly*," while text and audio contain irrelevant attributes. This setting tests per-modality reasoning and robustness to irrelevant signals.

Contradictory (⊕). Each modality leads to a different conclusion (Fig. 8). For example, text implies "*Erin is clean*", vision implies "*Bob is purple*", and audio implies "*Erin is tasty*". This interaction reveals the model's default preference when faced with conflicting evidence across modalities.

Complementary (∧). Each modality contributes a fact required for a conjunctive rule (Fig. 9). For instance, "*Erin is friendly*", "*Erin is purple*", and "*Erin is red*", are distributed across modalities, and the rule "*If a person is friendly and purple and red, then the person is clean.*" must be applied. This interaction evaluates the ability to integrate information across modalities for multi-source reasoning.

3 Evaluation of Modality Interactions: Pros and Cons

We now present a systematic evaluation of MLLMs across the six canonical interaction types. This section first outlines our experimental setup, then investigates when multiple modalities *help* or *hurt* reasoning, and finally synthesizes key bottlenecks revealed by the results.

3.1 Preparation

We begin by describing the models, prompting strategy, decoding procedure, and evaluation metric that together form a controlled testbed for analyzing modality interactions.

Models. We evaluate four recent open-source MLLMs that support at least three input modalities and generate text outputs. *Baichuan-Omni-1.5d (7B)* (Li et al., 2025c, Baichuan) is designed for efficient and balanced multimodal reasoning. *Qwen2.5-Omni (7B)* (Xu et al., 2025, Qwen) is a state-of-the-art model that handles text, vision, audio, and video, with streaming outputs, serving as a strong open baseline for multi-source reasoning. *MiniCPM-o-2.6 (8B)* (Yao et al., 2024, MiniCPM) processes text, vision, and audio, and is optimized for real-time multimodal streaming and on-device deployment. *Phi-4 Multimodal (5.6B)* (Abouelenin et al., 2025, Phi4) extends the Phi family to vision and audio, emphasizing compactness and efficiency over scale.[3] This suite covers both high-capacity and lightweight systems, enabling comparison of modality interaction across diverse architectures.

Prompt design and decoding. We use a unified prompt format across all models to ensure fair comparison. Each prompt includes a system instruction, a set of fact blocks in random modality order (text, vision, audio), a series of textual reasoning rules, and a four-way multiple-choice question (Fig. 1c). To encourage step-by-step reasoning, we insert concise CoT hints while minimizing behavioral interference. Decoding follows HuggingFace defaults with greedy sampling to produce stable outputs, from which the final answer is automatically extracted. All models are evaluated on the same synthetic dataset using identical prompt templates. More details are provided in App. B.3.

Evaluation metric. We report accuracy as the primary evaluation metric. We evaluate the reasoning performance by multiple-choice question answering format with four options, and the model's selected answer is automatically extracted from its output response. Since there are four options, the random guessing would yield a baseline accuracy of 25%. To ensure robustness, each experiment is conducted on 1,000 synthetic instances per condition.

[3]We refer to these models by shortened names in all experiments.

3.2 Do Multiple Modalities Help Reasoning?

One motivation for using MLLMs is to enhance reasoning by incorporating additional information from multiple modalities. But does adding useful input from another modality always help? To explore this, we evaluate three controlled interaction types designed to probe distinct modes of cross-modal benefit: redundancy, optionality, and composition. Results are summarized in Tab. 1, with full breakdowns in Apps. C.1 to C.3.

Setup. In *Equivalence*, the same decisive fact is redundantly placed in all modalities, testing whether repetition reinforces reasoning. In *Alternative*, each modality contains a distinct but individually sufficient fact for solving the problem, allowing multiple independent reasoning paths. In *Entailment*, a reasoning chain (A $\rightarrow$ B $\rightarrow$ C $\rightarrow$ Answer) is split across modalities, requiring integration of cross-modal premises for successful inference. To assess the added value of multimodal input, we compare each multimodal setting to its unimodal baselines, where only one decisive fact is in one modality (text, vision, or audio) and the others are omitted. From this comparison we can directly analyze the additional value brought by additional information in extra modalities.

Table 1: **Does Multimodality Help Reasoning?** Accuracy (%) and performance deltas (Δ) relative to unimodal baselines across three interaction types: *Equivalence* (redundant facts across modalities), *Alternative* (independent reasoning paths), and *Entailment* (multi-hop chains split across modalities, with final-step facts in V/A/T respectively). *Alternative* settings slightly boost performance, *Equivalence* yields marginal decrease (compared to text), while *Entailment* causes notable accuracy drops on reasoning.

Accuracy (%)	**Multimodal ($\equiv$, $\vee$, $\rightarrow$)**				
	Equivalence$_{\Delta V,\Delta A,\Delta T}$	**Alternative**$_{\Delta V,\Delta A,\Delta T}$	**Entailment**: $\mathbf{V}_{\Delta V}$	$\mathbf{A}_{\Delta A}$	$\mathbf{T}_{\Delta T}$
Baichuan	$84.8_{5.4\uparrow, 9.8\uparrow, 11.1\downarrow}$	$97.6_{19.6\uparrow, 17.8\uparrow, 0.3\uparrow}$	$79.5_{2.0\downarrow}$	$75.6_{6.4\downarrow}$	$80.7_{13.6\downarrow}$
Qwen	$98.9_{2.6\uparrow, 4.5\uparrow, 0.9\uparrow}$	$100.0_{3.7\uparrow, 6.1\uparrow, 2.6\uparrow}$	$78.4_{15.7\downarrow}$	$86.6_{8.2\downarrow}$	$83.9_{12.8\downarrow}$
MiniCPM	$94.8_{5.4\uparrow, 5.2\uparrow, 0.2\downarrow}$	$99.1_{7.1\uparrow, 8.0\uparrow, 2.9\uparrow}$	$81.8_{11.4\downarrow}$	$80.0_{12.0\downarrow}$	$88.4_{6.8\downarrow}$
Phi4	$84.1_{25.3\uparrow, 23.9\uparrow, 12.5\downarrow}$	$97.9_{20.3\uparrow, 26.3\uparrow, 1.0\uparrow}$	$73.0_{2.2\downarrow}$	$69.3_{0.7\downarrow}$	$79.7_{18.0\downarrow}$
Average	$90.7_{9.7\uparrow, 10.9\uparrow, 5.7\downarrow}$	$98.7_{12.7\uparrow, 14.8\uparrow, 1.7\uparrow}$	$78.2_{7.8\downarrow}$	$77.9_{7.1\downarrow}$	$83.2_{12.8\downarrow}$

Findings. In *Equivalence*, models show marginal gains when the decisive fact is in vision (+9.7%) or audio (+10.9%), but performance drops when the fact is already in text (-5.7%). This suggests that redundancy is only helpful when the original modality is weak. In *Alternative*, consistent improvements are observed across all modalities (+12.7% vision, +14.8% audio, +1.7% text), indicating that semantically independent reasoning paths are successfully leveraged. However, *Entailment* leads to substantial drops in accuracy across all modalities (-7.8% vision, -7.1% audio, -12.8% text), highlighting the difficulty of cross-modal multi-hop composition.

Observation 1. Multimodal input improves reasoning only when it contributes additional, semantically independent reasoning paths. In contrast, redundant information provides little benefit, particularly when a strong modality (text) is already sufficient, and distributing multi-step reasoning chains across modalities often reduces accuracy. These results suggest that *the core bottleneck in multimodal reasoning lies not in recognizing facts*, since individual modalities suffice in many cases.

3.3 Do Multiple Modalities Hurt Reasoning?

While certain forms of multimodal input can aid reasoning (§ 3.2), adding modalities could also introduce errors. In this section, we evaluate three controlled settings: *Independence*, *Contradictory*, and *Complementary*: to identify specific failure modes where multiple modalities degrade reasoning.

3.3.1 Modality Performance Bias: Independence

We first explore whether models exhibit consistent reasoning *performance* across modalities.

Setup. In the *Independence* setting, a decisive fact appears in one modality (text, vision, or audio), while the remaining modalities contain only distractors. We compare multimodal reasoning, where

facts are distributed across modalities, to unimodal baselines, where all facts (both decisive and distracting) are presented within a single modality.

Findings. As shown in Tab. 2 (full results can be found in App. C.4), models perform best in the text-only condition (94.45% on average), but accuracy drops sharply to 70.29% when facts are distributed across modalities. This is well below text-only performance but above vision-only or audio-only baselines, confirming that weaker modalities introduce noise when combined with stronger ones.

Table 2: **Performance on the *Independence* interaction.** Each instance includes one decisive fact placed in a single modality, while the others contain distractors. Multimodal reasoning accuracy falls between the best (text) and worst (vision) unimodal conditions, suggesting that modality inconsistency introduces error when aggregating information across modalities.

Accuracy (%)	**Unimodal**			**Multimodal** $(\emptyset)_{\Delta\mathbf{V},\Delta\mathbf{A},\Delta\mathbf{T}}$
	V	**A**	**T**	
Baichuan	60.2	72.0	94.8	$67.6_{7.4\uparrow, 4.4\downarrow, 27.2\downarrow}$
Qwen	73.3	94.3	95.5	$75.2_{1.9\uparrow, 19.1\downarrow, 20.3\downarrow}$
MiniCPM	77.6	83.7	91.2	$78.7_{1.1\uparrow, 5.0\downarrow, 12.5\downarrow}$
Phi4	49.9	48.9	96.3	$59.7_{9.8\uparrow, 10.8\uparrow, 36.6\downarrow}$
Average	65.3	74.7	94.5	$70.3_{5.0\uparrow, 4.4\downarrow, 24.2\downarrow}$

Observation 2. Unequal reasoning capabilities across modalities, what we refer to as *performance bias*, contribute significantly to degraded multimodal reasoning. When weaker modalities are added, they can dilute or confuse the signal from stronger ones, like text.

3.3.2 Modality Preference Bias: Contradictory

We next ask whether models exhibit internal *preferences* for certain modalities.

Setup. In the *Contradictory* setting, each modality provides a distinct and individually sufficient reasoning path, but the answer options are mutually exclusive.[4] This setup exposes which modality a model relies on when conflicting information is presented. Crucially, preference here refers to *selection behavior under conflict*, not standalone performance.

Table 3: Performance on the **Contradictory** interaction, where each modality leads to a different answer. Models show clear modality preferences, highlighting inconsistent reliance on input sources.

Answer Ratio (%)	**Multimodal** $(\oplus)$		
	V	**A**	**T**
Baichuan	49.0	14.9	33.7
Qwen	17.2	44.6	37.6
MiniCPM	22.6	27.2	49.0
Phi4	31.9	19.1	46.1

Results. As shown in Tab. 3, models display clear preference patterns: Baichuan favors vision-based answers (49.0%), Qwen tends toward audio (44.6%), and both MiniCPM and Phi4 prefer text (49.0% and 46.1%). These choices are often misaligned with the models' unimodal strengths, suggesting implicit biases in modality selection under conflicting input.

Observation 3. In addition to performance bias, MLLMs also suffer from *preference bias*: when modalities conflict, models favor certain modalities instead of strong modalities, often inconsistently with the actual performance. This misalignment introduces further risk in multimodal reasoning.

3.3.3 Modality Fusion Bias: Complementary

Finally, we examine whether models can *fuse* complementary evidence across modalities when each input is necessary for inference.

Setup. In *Complementary* setting, each modality contains one of three facts that are jointly required to solve the reasoning task. Unlike prior settings with a single decisive fact, here all facts must be composed across modalities. We compare this condition to unimodal baselines where all three facts are provided within a one modality.

[4]Note that one of the four answer options is always incorrect by design, so the sum of selection ratios across modalities does not equal 100%.

Results. In Tab. 4, all models perform *worse* in the multimodal setup than in any unimodal condition, even when confined to weaker modalities like vision. If biased performance is the only issue, multimodal accuracy should lie between the best and worst unimodal conditions. Instead, distributing complementary facts across modalities introduces a new failure mode: models are unable to compose multiple weak signals into a coherent reasoning chain.

Table 4: **Performance on the *Complementary* interaction.** Each modality provides one necessary fact, requiring to integrate all three to get the answer. The multimodal reasoning accuracy is lower than any unimodal condition, indicating that performance degradation stems not only from modality inconsistency but also from a true cross-modal composition bottleneck.

Accuracy (%)	**Unimodal**			**Multimodal** $(\wedge)_{\Delta\mathbf{V},\Delta\mathbf{A},\Delta\mathbf{T}}$
	V	**A**	**T**	
Baichuan	50.5	59.4	87.7	$30.2_{20.3\downarrow, 29.2\downarrow, 57.5\downarrow}$
Qwen	87.5	98.8	98.8	$49.9_{37.6\downarrow, 48.9\downarrow, 48.9\downarrow}$
MiniCPM	74.8	89.3	92.4	$48.8_{26.0\downarrow, 40.5\downarrow, 43.6\downarrow}$
Phi4	80.0	82.2	99.6	$79.1_{0.9\downarrow, 3.1\downarrow, 20.5\downarrow}$
Average	73.2	82.4	94.6	$52.0_{21.2\downarrow, 30.4\downarrow, 42.6\downarrow}$

Observation 4. MLLMs struggle to integrate complementary information across modalities, even when all inputs are individually comprehensible. This reveals that beyond biased performance and preference, there is a third failure model in multimodal reasoning: *fusion bias*.

3.4 From Observations to Bottlenecks

The preceding evaluations reveal consistent patterns in the way models handle multimodal reasoning. We now synthesize these findings to identify the underlying bottlenecks.

Observation 5. Across all settings, the best performance consistently comes from the text-only baseline, often approaching near-perfect accuracy. This shows that models can already perform logical reasoning reliably when inputs are centralized in a single strong modality. Combined with **Observation 1** (models can recognize facts across modalities), this indicates that the core weakness lies not in perception or reasoning in isolation, but in how these components are combined.

Bottleneck 1: Task Composition. The conjunction of **Observation 1** (strong recognition) and **Observation 5** (strong unimodal reasoning) points to a first core bottleneck: *task composition*. Models falter when recognition and reasoning must be performed jointly across modalities. That is, while they can detect facts and apply reasoning rules when each task is isolated, performance drops sharply when these steps must be integrated within a single inference pass.

Bottleneck 2: Multi-Source Fusion. **Observations 2–4** collectively point to a second, orthogonal failure point: the inability to *fuse* information from multiple modalities in a reliable and unbiased manner. First, performance bias (**Observation 2**) shows that weak modalities dilute reasoning when mixed with stronger ones. Second, preference bias (**Observation 3**) reveals that models often favor certain modalities under conflict, even when those modalities underperform in isolation. Third, fusion bias (**Observation 4**) shows that models fail to integrate complementary information spread across modalities, even when all inputs are individually comprehensible. These findings suggest that MLLMs lack robust internal mechanisms for selecting, weighting, and composing evidence from heterogeneous sources, leading to systematic failures in multimodal reasoning.

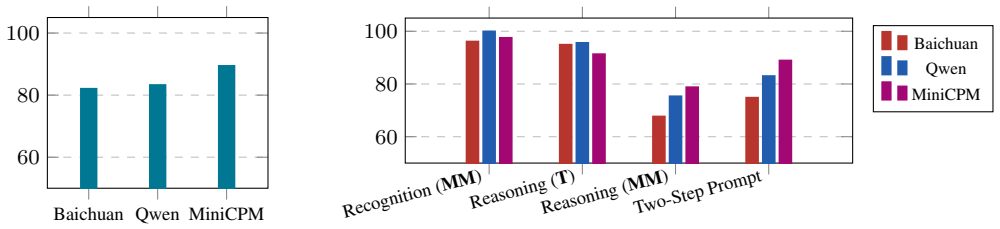
4 Interpretation: Bottlenecks of Multimodal Reasoning

We now investigate the underlying causes of multimodal reasoning failures by probing the internal mechanisms of MLLMs. Guided by the two key bottlenecks identified in § 3: *task composition* and *multi-source fusion*, we analyze how models represent modality and information usefulness, and whether targeted interventions can mitigate these bottlenecks.[5]

[5]In this section, we conduct our analysis under the *Independence* setting to avoid multi-hop confounds and ensure interpretability. We select three representative models, Baichuan, Qwen, and MiniCPM, based on their differing modality preferences identified in § 3.3.2, which favor vision, audio, and text, respectively.

4.1 Task Composition: Analysis and Improvement

We begin with **Bottleneck 1**, which concerns the model's inability to compose recognition and reasoning across modalities within a single inference step. While MLLMs can perceive facts from different modalities and reason over them in isolation, their performance degrades when these abilities must be integrated. We investigate this bottleneck by analyzing internal attention patterns and exploring whether prompting strategies can mitigate the failure.



(a) Probe Acc. (%) **(b)** Reasoning Accuracy (%)

Figure 2: **Attention probing and reasoning performance.** (a) Modality probing for information usefulness shows moderate accuracy, suggesting models cannot clearly distinguish useful from distractor facts. (b) Although models excel in fact recognition and text-only reasoning, their performance drops significantly on multimodal reasoning, indicating that the key limitation lies in composing recognition and reasoning across modalities.

Probing Information Usefulness. We first assess whether models can internally distinguish useful facts from distractors. A linear probe is trained on decoder attention distributions: for each modality, we compute mean attention over all generated tokens and use these vectors to classify whether each fact is relevant for reasoning (more probing details are in App. B.2). Results (Fig. 2a) show that the attention patterns do not reliably signal semantic usefulness, indicating that the models struggle to prioritize decision-relevant content based on attention alone.

Interpretation Results. To isolate where failures occurs, we conduct two controlled diagnostics: (1) a formal *recognition test*, where models identify facts across modalities without requiring additional reasoning (see Fig. 10), and (2) a *two-step prompting* setup, where recognition and reasoning are separated across prompts (see Fig. 11). As shown in Fig. 2b, models perform near-perfectly on recognition and maintain high accuracy on unimodal reasoning. However, when both recognition and reasoning are combined in a single multimodal prompt, accuracy drops substantially, confirming that the core failure lies in the integration of these two capabilities.

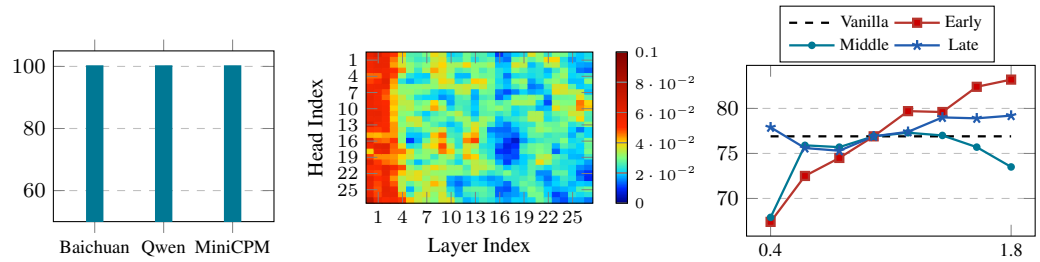
Improving Composition via Two-Step Prompting. In the two-step prompting strategy, models first extracting all facts, then reason over them, substantially improving performance across all models. This shows that the failure arises not from deficiencies in perception or reasoning alone, but from their joint composition within a single inference step. By decoupling these processes and providing a more explicit recognition goal, the task-composition bottleneck is effectively alleviated.

Takeaway. These findings highlight a core weakness in current MLLMs: despite strong perception and reasoning abilities in isolation, they lack mechanisms to integrate these steps across modalities. This reflects a broader limitation in training objectives, which emphasize shallow alignment rather than compositional inference. Prompt-level task decomposition offers a simple yet effective remedy.

4.2 Modality Fusion: Analysis and Improvement

We now address **Bottleneck 2**, which concerns the model's difficulty in fusing information across modalities. Even when individual facts are recognized correctly, reasoning often fails due to biased or ineffective modality integration. To better understand this bottleneck, we analyze how modality identity is internally represented and whether early fusion contributes to systematic errors.

Probing Modality Identity. We assess whether modality type (text, vision, audio) is preserved in the model's internal representations. A logistic regression classifier is trained on attention-derived features: for each input fact, similarly, we compute average attention from all generated tokens and



(a) Probe Acc. (%) **(b)** Linear probe weights for Qwen **(c)** Reasoning Acc. (%) for Qwen

Figure 3: **Modality probing based on attention patterns.** (a) All models achieve perfect probe accuracy in predicting the modality using attention patterns. (b) For Qwen, linear probe weights show that modality information is primarily captured in the first four layers. (c) Attention manipulation in different 4 layers (by adjusting head temperature from 0.4 to 1.8), where performance significantly improves in the early 4 layers.

flatten this into a feature vector. The classifier predicts the modality of each fact. As shown in Fig. 3a, modality identity is perfectly recoverable, indicating that even after fusion, the model maintains a strong internal signal of the input modality.

Interpretation Results. To locate where modality fusion occurs, we visualize layer-wise probe weights. Fig. 3b shows that the first four decoder layers carry the strongest modality signal, suggesting that fusion predominantly occurs early in the language module. Beyond this point, modalities appear to be processed more uniformly.

Improving Fusion via Attention Manipulation. Motivated by this early-fusion pattern, we modify attention behavior by adjusting the softmax temperature from 0.4 to 1.8 in the first four decoder layers (Early), with the default set to 1.0. As shown in Fig. 3c, this simple intervention of increasing the temperature for early layers yields significant improvements in reasoning accuracy by encouraging more balanced attention across modalities. In contrast, adjusting the temperature in middle or late layers has little effect, supporting the causal role of early fusion in downstream reasoning outcomes.

Takeaways. These findings confirm that while modality identity is well preserved, early-stage fusion introduces systematic biases that impair reasoning. A lightweight causal intervention, reshaping early attention distributions, can significantly enhance multimodal integration, highlighting the importance of fusion dynamics over perceptual bottlenecks.

Summary of Interpretation Findings. Across above interpretation analyses, we find that failures in multimodal reasoning stem not from deficiencies in perception or unimodal reasoning, but from weak *compositional integration* and *cross-modal fusion*. Although MLLMs retain modality-specific signals and can recognize facts reliably, they often default to shallow alignment behavior rather than selective integration. Without architectural biases or training objectives that explicitly encourage multimodal composition, these models remain brittle in complex reasoning tasks.

5 Conclusion

This work presents a systematic study of how MLLMs integrate information across modalities for logical reasoning. We introduce a logic-driven evaluation framework with six canonical interaction types, enabling controlled analysis of when additional modalities help or hinder reasoning. Our results reveal a consistent pattern: modalities help only when they contribute independent, sufficient reasoning paths, while redundancy or cross-modal chaining often degrades performance. Text-only baselines already approach ceiling accuracy, underscoring that the key bottleneck is not perception but integration. Through probing and causal interventions, we identified two core bottlenecks: a *task-composition bottleneck*, where recognition and reasoning cannot be reliably combined in a single inference step, and a *fusion bottleneck*, where early-layer integration introduces modality bias. Simple remedies such as two-step prompting and attention temperature adjustments alleviate these issues, highlighting the importance of designing models and objectives that explicitly support evidence selection and unbiased fusion. We hope our framework and findings inspire future work toward composition-aware training and architecture choices that transform added modalities from sources of interference into assets for reasoning.

ETHICS STATEMENT

This research does not involve human participants, private or sensitive data, or applications with foreseeable negative societal impact. All datasets employed are publicly available and widely used within the vision–language and reasoning research communities. We adhere to standard best practices in data handling, model evaluation, and reporting, and our study complies fully with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have taken care to make our work reproducible. The paper and appendix provide detailed descriptions of the models, datasets, and experimental setups. Code, data generation scripts, and evaluation protocols are included in the supplementary material and will be released publicly upon publication to facilitate replication and further research.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Jun-Kun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *CoRR*, abs/2503.01743, 2025. doi: 10.48550/ARXIV.2503.01743. URL https://doi.org/10.48550/arXiv.2503.01743.

Jing Bi, Junjia Guo, Susan Liang, Guangyu Sun, Luchuan Song, Yunlong Tang, Jinxi He, Jiarui Wu, Ali Vosoughi, Chen Chen, and Chenliang Xu. VERIFY: A benchmark of visual explanation and reasoning for investigating multimodal reasoning fidelity. *CoRR*, abs/2503.11557, 2025. doi: 10.48550/ARXIV.2503.11557. URL https://doi.org/10.48550/arXiv.2503.11557.

Fuqing Bie, Shiyu Huang, Xijia Tao, Zhiqin Fang, Leyi Pan, Junzhe Chen, Min Ren, Liuyu Xiang, and Zhaofeng He. Omniplay: Benchmarking omni-modal models on omni-modal game playing, 2025. URL https://arxiv.org/abs/2508.04361.

Herbert H. Clark and Susan E. Brennan. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley (eds.), *Perspectives on socially shared cognition*, pp. 127–149. American Psychological Association, 1991. doi: 10.1037/10096-006. URL https://doi.org/10.1037/10096-006.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pp. 3882–3890. ijcai.org, 2020. doi: 10.24963/IJCAI.2020/537. URL https://doi.org/10.24963/ijcai.2020/537.

Bruce Coburn, Jiangpeng He, Megan E. Rollo, Satvinder S. Dhaliwal, Deborah A. Kerr, and Fengqing Zhu. Evaluating large multimodal models for nutrition analysis: A benchmark enriched with contextual metadata. *CoRR*, abs/2507.07048, 2025. doi: 10.48550/ARXIV.2507.07048. URL https://doi.org/10.48550/arXiv.2507.07048.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023. doi: 10.48550/ARXIV.2306.13394. URL https://doi.org/10.48550/arXiv.2306.13394.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24108–24118. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02245. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion

in large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14375–14385. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01363. URL https://doi.org/10.1109/CVPR52733.2024.01363.

Himanshu Gupta, Shreyas Verma, Ujjwala Anantheswaran, Kevin Scaria, Mihir Parmar, Swaroop Mishra, and Chitta Baral. Polymath: A challenging multi-modal mathematical reasoning benchmark. *CoRR*, abs/2410.14702, 2024. doi: 10.48550/ARXIV.2410.14702. URL https://doi.org/10.48550/arXiv.2410.14702.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? EMMA: an enhanced multimodal reasoning benchmark. *CoRR*, abs/2501.05444, 2025. doi: 10.48550/ARXIV.2501.05444. URL https://doi.org/10.48550/arXiv.2501.05444.

Abdelhamid Haouhat, Slimane Bellaouar, Attia Nehar, and Hadda Cherroun. Modality influence in multimodal machine learning. *CoRR*, abs/2306.06476, 2023. doi: 10.48550/ARXIV.2306.06476. URL https://doi.org/10.48550/arXiv.2306.06476.

Yixiao He, Haifeng Sun, Pengfei Ren, Jingyu Wang, Huazheng Wang, Qi Qi, Zirui Zhuang, and Jing Wang. Evaluating and mitigating object hallucination in large vision-language models: Can they still see removed objects? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 6841–6858. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.349. URL https://doi.org/10.18653/v1/2025.naacl-long.349.

Yifan Hou, Buse Giledereli, Yilei Tu, and Mrinmaya Sachan. Do vision-language models really understand visual language?, 2025. URL https://arxiv.org/abs/2410.00193.

Songtao Jiang, Yan Zhang, Ruizhe Chen, Tianxiang Hu, Yeying Jin, Qinglin He, Yang Feng, Jian Wu, and Zuozhu Liu. Modality-fair preference optimization for trustworthy mllm alignment, 2025. URL https://arxiv.org/abs/2410.15334.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025, 2025a. URL https://openreview.net/forum?id=zKv8qULV6n.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *CoRR*, abs/2307.16125, 2023. doi: 10.48550/ARXIV.2307.16125. URL https://doi.org/10.48550/arXiv.2307.16125.

Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. *CoRR*, abs/2506.04633, 2025b. doi: 10.48550/ARXIV.2506.04633. URL https://doi.org/10.48550/arXiv.2506.04633.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, Yuran Wang, Bowen Ding, Wei Song, Xu Li, Yuqi Huo, Zheng Liang, Shusen Zhang, Xin Wu, Shuai Zhao, Linchu Xiong, Yozhen Wu, Jiahui Ye, Wenhao Lu, Bowen Li, Yan Zhang, Yaqi Zhou, Xin Chen, Lei Su, Hongda Zhang, Fuzhong Chen, Xuezhen Dong, Na Nie, Zhiying Wu, Bin Xiao, Ting Li, Shunya Dang, Ping Zhang, Yijia Sun, Jincheng Wu, Jinjie Yang, Xionghai Lin, Zhi Ma, Kegeng Wu, Jia Li, Aiyuan Yang, Hui Liu, Jianqiang Zhang, Xiaoxi Chen, Guangwei Ai, Wentao Zhang, Yicong Chen, Xiaoqin Huang, Kun Li, Wenjing Luo, Yifei Duan, Lingling Zhu, Ran Xiao, Zhe Su, Jiani Pu, Dian Wang, Xu Jia, Tianyu Zhang, Mengyu Ai, Mang Wang, Yujing Qiao, Lei Zhang, Yanjun Shen, Fan Yang, Miao Zhen, Yijie Zhou, Mingyang Chen, Fei Li, Chenzheng Zhu, Keer Lu, Yaqi Zhao, Hao Liang, Youquan Li, Yanzhao Qin, Linzhuang Sun, Jianhua Xu, Haoze Sun, Mingan Lin, Zenan Zhou, and Weipeng Chen. Baichuan-omni-1.5 technical report. *CoRR*, abs/2501.15368, 2025c. doi: 10.48550/ARXIV.2501.15368. URL https://doi.org/10.48550/arXiv.2501.15368.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. Omnibench: Towards the future of universal omni-language models. *CoRR*, abs/2409.15272, 2024. doi: 10.48550/ARXIV.2409.15272. URL https://doi.org/10.48550/arXiv.2409.15272.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: on pre-training for visual language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26679–26689. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02520. URL https://doi.org/10.1109/CVPR52733.2024.02520.

Ronghao Lin and Haifeng Hu. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Trans. Assoc. Comput. Linguistics*, 11:1686–1702, 2023. doi: 10.1162/TACL_A_00628. URL https://doi.org/10.1162/tacl_a_00628.

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 Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LVI*, volume 15114 of *Lecture Notes in Computer Science*, pp. 386–403. Springer, 2024a. doi: 10.1007/978-3-031-72992-8_22. URL https://doi.org/10.1007/978-3-031-72992-8_22.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pp. 216–233. Springer, 2024b. doi: 10.1007/978-3-031-72658-3_13. URL https://doi.org/10.1007/978-3-031-72658-3_13.

Richard E Mayer. Multimedia learning. In *Psychology of learning and motivation*, volume 41, pp. 85–139. Elsevier, 2002.

Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, Yangui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, Shuai Wang, and Kai Yu. A survey on speech large language models for understanding, 2025. URL https://arxiv.org/abs/2410.18908.

Shaina Raza, Ashmal Vayani, Aditya Jain, Aravind Narayanan, Vahid Reza Khazaie, Syed Raza Bashir, Elham Dolatabadi, Gias Uddin, Christos Emmanouilidis, Rizwan Qureshi, and Mubarak Shah. Vldbench evaluating multimodal disinformation with regulatory alignment, 2025. URL https://arxiv.org/abs/2502.11361.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha

Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530, 2024. doi: 10.48550/ARXIV.2403.05530. URL https://doi.org/10.48550/arXiv.2403.05530.

Yusuf Shihata. Gated recursive fusion: A stateful approach to scalable multimodal transformers. *CoRR*, abs/2507.02985, 2025. doi: 10.48550/ARXIV.2507.02985. URL https://doi.org/10.48550/arXiv.2507.02985.

Anoop K. Sinha, Chinmay Kulkarni, and Alex Olwal. Levels of multimodal interaction. In Hayley Hung, Catharine Oertel, Mohammad Soleymani, Theodora Chaspari, Hamdi Dibeklioglu, Jainendra Shukla, and Khiet P. Truong (eds.), *Companion Proceedings of the 26th International Conference on Multimodal Interaction, ICMI Companion 2024, San Jose, Costa Rica, November 4-8, 2024*, pp. 51–55. ACM, 2024. doi: 10.1145/3686215.3690153. URL https://doi.org/10.1145/3686215.3690153.

Christopher Thomas, Yipeng Zhang, and Shih-Fu Chang. Fine-grained visual entailment. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pp. 398–416. Springer, 2022. doi: 10.1007/978-3-031-20059-5_23. URL https://doi.org/10.1007/978-3-031-20059-5_23.

Shakti N. Wadekar, Abhishek Chaurasia, Aman Chadha, and Eugenio Culurciello. The evolution of multimodal model architectures. *CoRR*, abs/2405.17927, 2024. doi: 10.48550/ARXIV.2405.17927. URL https://doi.org/10.48550/arXiv.2405.17927.

Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. When language overrules: Revealing text dominance in multimodal large language models, 2025. URL https://arxiv.org/abs/2508.10552.

Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *CoRR*, abs/2410.11190, 2024. doi: 10.48550/ARXIV.2410.11190. URL https://doi.org/10.48550/arXiv.2410.11190.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215, 2025. doi: 10.48550/ARXIV.2503.20215. URL https://doi.org/10.48550/arXiv.2503.20215.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Haofei Yu, Zhengyang Qi, Lawrence Jang, Russ Salakhutdinov, Louis-Philippe Morency, and Paul Pu Liang. Mmoe: Enhancing multimodal models with mixtures of multimodal interaction experts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 10006–10030. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.558. URL https://doi.org/10.18653/v1/2024.emnlp-main.558.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. RLAIF-V: open-source AI feedback leads to super GPT-4V trustworthiness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 19985–19995. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01861. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yu_RLAIF-V_Open-Source_AI_Feedback_Leads_to_Super_GPT-4V_Trustworthiness_CVPR_2025_paper.html.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9556–9567. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00913. URL https://doi.org/10.1109/CVPR52733.2024.00913.

Yu Zhang, Jinlong Ma, Yongshuai Hou, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. Evaluating and steering modality preferences in multimodal large language model. *CoRR*, abs/2505.20977, 2025. doi: 10.48550/ARXIV.2505.20977. URL https://doi.org/10.48550/arXiv.2505.20977.

Fei Zhao, Taotian Pang, Chunhui Li, Zhen Wu, Junjie Guo, Shangyu Xing, and Xinyu Dai. Aligngpt: Multi-modal large language models with adaptive alignment capability, 2024.

A Related Work

Multimodal Large Language Models (MLLMs). Recent MLLMs extend text-only LLMs to handle images, audio, and often video, with two broad design trends. On the end-to-end omni side, Qwen2.5-Omni processes text–vision–audio–video with streaming IO via a block-wise encoder stack and a Thinker–Talker architecture for joint text/speech generation, achieving state-of-the-art results on omni benchmarks (Xu et al., 2025). MiniCPM-o-2.6 follows a similarly integrated pipeline (SigLIP + Whisper + ChatTTS + Qwen2.5 backbone) to enable real-time speech and multimodal live streaming on resource-constrained devices (Yu et al., 2025). Baichuan-Omni proposes a two-stage scheme, multimodal alignment, then multitask fine-tuning, to support concurrent inputs of image, audio, video, and text in an open 7B model (Li et al., 2025c). In contrast, adapter-based approaches extend a compact LLM with modality-specific LoRA routers (e.g., Phi-4 Multimodal), delivering strong vision–audio performance while retaining efficiency (Abouelenin et al., 2025). Beyond open models, frontier systems like Gemini emphasize extremely long multimodal context across documents, video, and audio (Reid et al., 2024). Complementary open MLLM lines (LLaVA-OneVision (Li et al., 2025a), VILA-1.5 (Lin et al., 2024)) focus on unifying single-image, multi-image, and video scenarios and scaling families across 3B–40B parameters, respectively. Finally, efforts like Mini-Omni2 target GPT-4o-style, real-time visual–audio assistants, underscoring the field's push toward unified, low-latency multimodal interaction (Xie & Wu, 2024). Recent surveys also review progress in Speech LLMs, formally defining speech understanding and analyzing Speech-LLM architectures, training, and evaluation through a structured taxonomy (Peng et al., 2025).

MLLM Evaluation. A rapidly growing ecosystem of benchmarks now probes different facets of the MLLM ability. General-purpose suites such as MMBench (Liu et al., 2024b), MME (Fu et al., 2023), and SEED-Bench (Li et al., 2023) emphasize breadth and stable, objective MCQ evaluation across perception and language understanding, including text-rich images and video. Reasoning-centric datasets like MMMU push models toward college-level, multi-disciplinary problem solving with heterogeneous visual artifacts (charts, diagrams, tables), exposing large gaps in expert-level multimodal reasoning (Yue et al., 2024). Complementing capability scores, robustness-focused evaluations, including POPE (He et al., 2025) for object hallucination and HallusionBench (Guan et al., 2024) for language-vs-vision conflicts, and visual illusions-diagnose systematic failure modes and modality biases that broad benchmarks can obscure. Beyond static images, Video-MME (Fu et al., 2025) targets temporal understanding, while MM-SafetyBench (Liu et al., 2024a) stress-tests safety in multimodal settings. These lines reveal a pattern: While breadth benchmarks track steady gains, targeted diagnostics consistently uncover modality dominance, hallucination, and fusion brittleness, motivating frameworks (like ours) that isolate and measure information interaction in reasoning.

Recognition vs. Reasoning. While many benchmarks and papers demonstrate that MLLMs are adept at recognizing objects, attributes, diagrams, or patterns, few dissect how well these models reason with those recognized elements, especially when evidence must be drawn from multiple modalities. Benchmarks like VERIFY (Bi et al., 2025) highlight that perception (recognition of visual input) is often less challenging for MLLMs than reasoning, models often fail when inference, explanation, or abstract relationships are required. STARE (Li et al., 2025b) similarly reveals that while simple spatial or 2D transformations can be handled, tasks that need multistep visual simulation or 3D understanding are far more error-prone. POLYMATH (Gupta et al., 2024) and EMMA (Hao et al., 2025) further push into domains (math, diagrams, cross-modal reasoning) where recognition alone is insufficient; these tasks expose gaps when models must combine or interpret recognized information rather than just identify it. Our work builds on this line by explicitly separating and quantifying fact recognition, modality recognition, and reasoning performance, and by introducing interaction types (e.g., independence, entailment, complementary) to isolate how modalities help or hurt when reasoning is required.

Information Interaction and Logical Relations. A number of studies have implicitly or explicitly considered how information from multiple modalities interacts, especially in logical or semantic reasoning contexts. For example, Thomas et al. (2022) asks whether textual hypotheses entail, contradict, or are neutral with respect to images, and breaks down hypothesis claims, which correspond roughly to "entailment" and "contradiction" in our framework. Meanwhile, Sinha et al. (2024) proposes a qualitative taxonomy that includes redundancy and synergy between modalities, similar to our

"equivalence" and "complementary" types, but without formal logical rules or controlled tasks to isolate their effects. The recent MMOE work (Yu et al., 2024) directly trains separate experts for redundancy, uniqueness, and synergy, which is closely aligned with some of our interaction types (independent, alternative, complementary), yet still coarser. Architectural analyses (Wadekar et al., 2024) show different fusion strategies (early vs late vs cross-attention) that implement interactions implicitly rather than measuring them via logical operators. Our work builds on and extends this prior art by defining six fine-grained logical interaction types (independent, equivalence, alternative, entailment, complementary, contradictory) and embedding them into synthetic reasoning tasks with rules, facts, and contrastive answer options. This allows us to evaluate not just whether models fuse modalities, but also how and when different interaction patterns boost or degrade reasoning performance.

Fusion Mechanisms and Modality Dominance. Research on fusion architectures and modality dominance has rapidly advanced, revealing how design choices often tip the scale in favor of one modality, typically text. For instance, Wu et al. (2025) introduces quantitative metrics to measure how strongly models depend on text, showing that fusion architectures and token redundancy play key roles in producing text dominance. Similarly, Zhang et al. (2025) analyze modality dominance in MLLMs and attribute it to imbalances in scaling, alignment, and representation, showing that text often overwhelms other modalities not only due to token abundance, but also architectural biases. Their controlled ablations further highlight that dominance emerges from systemic training choices rather than dataset artifacts. Meanwhile, Haouhat et al. (2023) uses masked-modality ablation across tasks (sentiment, emotion, disease detection) to confirm that, in many settings, non-text modalities add little when text is present. Architectural solutions have been proposed: Shihata (2025) implements gating and sequential fusion to control how each modality contributes. Lin & Hu (2023) further show that handling missing modalities and aligning feature/geometric spaces can reduce dominance bias. Our work links to this by not only observing modality dominance in reasoning tasks, but also by dissecting when and why dominance arises via logical interactions and attention probing. We contribute by explicitly measuring both performance (modality competence) and preference (which modality is used when conflicting or combined), under controlled logical reasoning settings.

B Supplementary Setup

B.1 Fact Construction

We construct facts and rules following the format of Clark et al. (2020), using the data generation code from Liang et al. (2023).[6] Below we describe the details of fact generation and rule generation.

Fact generation. Each fact consists of a *subject*, a *predicate*, and an *object*. The predicate is always the copula "is"; variability comes from the choice of subject and object. To keep the setting interpretable and avoid conflicts with commonsense priors, we prevent antonyms or synonyms from appearing together within the same instance. Details are as follows:

- **Subjects.** A subject is randomly sampled from three categories:
  - *Persons* (13 names): Alice, Bob, Carol, Dan, Erin, Frank, George, Harry, Iris, Jack, Kevin, Lance, Miller.
  - *Animals* (14 types): dog, cat, rabbit, mouse, tiger, lion, bear, squirrel, cow, panda, hedgehog, elephant, giraffe, hippo.
  - *Fruits* (15 types): apple, banana, orange, grape, strawberry, blueberry, watermelon, pineapple, mango, peach, cherry, pear, kiwi, lemon, plum.
- **Predicate.** Always "is" (e.g., "Bob *is* curious").
- **Objects.** Objects are adjective attributes describing the subject. We use a pool of 34 attributes: young, soft, scary, hot, smart, clean, beautiful, red, blue, green, purple, boring, strong, happy, round, big, noisy, fast, sticky, bouncy, spiky, furry, bright, shiny, magical, striped, spotted, tasty, juicy, toxic, friendly, curious, loud, sleepy.

[6] The code is from this GitHub repository.

Rule generation. Rules are constructed following Liang et al. (2023), with a minor modification for brevity:

- If the rule involves only a single attribute, we convert it into a compact nominal form to save tokens. Example: Original: "If a cow is weak, then the cow is small." Transformed: "Weak cow is small."
- If the rule involves multiple attributes, we retain the full "if–then" format for clarity. Example: "If a person is smart and young, then the person is curious."

Image & audio generation. For the multimodal reasoning experiments, we require factual information in both visual and auditory formats. To generate images, we employ the Graphviz toolkit, which effectively converts structured data into clear, labeled diagrams suitable for model interpretation. For audio synthesis, we utilize CosyVoice 2 (Du et al. (2024)), one of the top Text-to-Speech (TTS) model renowned for its high consistency in timbre, achieving human-parity synthesis quality. Since the audio is solely intended to provide semantic information for subsequent logical reasoning within the MLLM, and we are not investigating the model's advanced audio understanding capabilities, we prioritize accuracy over variability in the synthesized speech. Therefore, we use CosyVoice 2's default settings to ensure the speech is clear and precise, without introducing unnecessary characteristics that could potentially interfere with the reasoning task.

B.2 Experiment Setup

LLM inference setup. During evaluation, all models are run in float16 precision. For models that support audio output, we disable this feature and only generate text. The maximum number of generated tokens is set to 1024, which is typically sufficient for the model to produce a complete response, as long as the model does not exhibit output repetition. To ensure coherent and consistent output, we set 'do_sample=False' for all models, which is also the default setting for these models and considered best practice.

Linear probing setup. To construct the attention feature as the input of the probe, we extract and pool attentions. Specifically, for one data sample, the attention matrix shape should be $N \times L \times H \times O$. Here, N is the number of tokens representing one fact (in one modality); L is the number of layers in the model; H is the number of attention head in the model; and O is the number of generated tokens of the response for the model. Since O and N could vary in samples (or facts), we take the mean pooling in term of these two dimensions. The final attention feature for each sample is shaped as $L \times H$. We flatten the feature as one attention feature for one fact in one sample as the input for the probe classifier.

For the linear probe classifier, we use a 5-fold cross-validation setup with GroupKFold to evaluate the models. The models are trained with Logistic Regression using a regularization strength of C=1.0, an l2 penalty, and class_weight="balanced" to handle class imbalances. The feature vectors are normalized using the standard score method, which standardizes the features to have zero mean and unit variance. We apply these configurations to the task at hand. These configurations ensure a robust and consistent evaluation across all models tested.

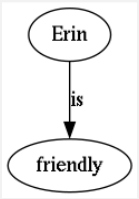
B.3 Prompt Templates

We list the prompt templates for all six types of interactions, the recognition task, and the two-step reasoning task. All models have the same input (template and data).

Prompt: input and output (Equivalence)

System: You are an assistant tasked with solving multiple-choice questions that require logical reasoning over the supplied knowledge diagrams. Use only the information explicitly given—do not rely on outside or commonsense knowledge. Read the question and given information, think step-by-step and answer the question. At the end of your answer, answer precisely in the format 'Answer: X' where X is the chosen letter A / B / C / D.

Input:



Erin is friendly

(Audio information: *Erin is friendly*)

Rules are as follows: Blue person is tasty. Red person is clean. If a person is smart or sleepy, then the person is curious. Friendly person is purple.
Which of the following options can be inferred based on the given facts and rules? A) Erin is curious. B) Erin is purple. C) Erin is tasty. D) Erin is clean.

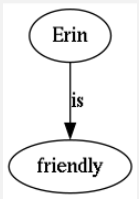
Output: To determine which option can be inferred based on the given facts and rules, let's analyze the information step by step:
1. Given Information: - Erin is friendly.
2. Rules: - Blue person is tasty. - Red person is clean. - If a person is smart or sleepy, then the person is curious. - Friendly person is purple.
3. Analysis: - From the rule ""Friendly person is purple,"" we can infer that since Erin is friendly, Erin must be purple.
4. Checking the Options:
- A) Erin is curious. The rule does not provide a direct link between being friendly and being curious. Therefore, we cannot infer this.
- B) Erin is purple. This directly follows from the rule "Friendly person is purple."
- C) Erin is tasty. The rule does not provide a direct link between being friendly and being tasty. Therefore, we cannot infer this.
- D) Erin is clean. The rule does not provide a direct link between being friendly and being clean. Therefore, we cannot infer this.
Based on the analysis, the only option that can be inferred is:
Answer: B

Figure 4: Prompt template and model (Qwen) output (Equivalence).

Prompt: input and output (Alternative)

System: You are an assistant tasked with solving multiple-choice questions that require logical reasoning over the supplied knowledge diagrams. Use only the information explicitly given—do not rely on outside or commonsense knowledge. Read the question and given information, think step-by-step and answer the question. At the end of your answer, answer precisely in the format 'Answer: X' where X is the chosen letter A / B / C / D.

Input:



Erin is purple

(Audio information: *Erin is red*)

Rules are as follows: Friendly person is clean. If a person is smart or sleepy, then the person is curious. Purple person is clean. Blue Erin is tasty. Spotted Erin is beautiful. Red person is clean.
Which of the following options can be inferred based on the given facts and rules? A) Erin is clean. B) Erin is tasty. C) Erin is beautiful. D) Erin is curious.

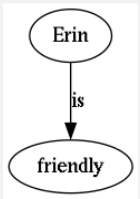
Output:
To determine which option can be inferred based on the given facts and rules, let's analyze the information step by step:
1. Given Information:
- Erin is purple.
- Erin is red.
- Erin is friendly.
2. Rules: - If a person is smart or sleepy, then the person is curious. - Blue Erin is tasty. - Spotted Erin is beautiful. - Friendly person is clean. - Purple person is clean. - Red person is clean.
3. Analysis: - Since Erin is purple and red, and the rules state that purple and red people are clean, we can infer that Erin is clean. - The rules do not provide information about Erin being smart, sleepy, curious, beautiful, or tasty based on the given facts.
4. Conclusion: - The only inference we can make from the given information is that Erin is clean.
Answer: A

Figure 5: Prompt template and model (Qwen) output (Alternative).

Prompt: input and output (Entailment)

System: You are an assistant tasked with solving multiple-choice questions that require logical reasoning over the supplied knowledge diagrams. Use only the information explicitly given—do not rely on outside or commonsense knowledge. Read the question and given information, think step-by-step and answer the question. At the end of your answer, answer precisely in the format 'Answer: X' where X is the chosen letter A / B / C / D.

Input:



Erin is bright

(Audio information: *Erin is bouncy*)

Rules are as follows: If a person is smart or sleepy, then the person is curious. Red person is clean. Blue person is tasty. Bright person is friendly. Friendly person is purple. Bouncy person is bright.
Which of the following options can be inferred based on the given facts and rules? A) Erin is curious. B) Erin is tasty. C) Erin is purple. D) Erin is clean.

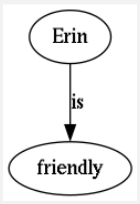
Output: Let's analyze the given information step by step:
1. Erin is bouncy.
2. Bouncy Erin is bright.
3. Erin is friendly.
4. Friendly person is purple.
5. Bright Erin is friendly.
From the information, we can deduce the following:
- Since Erin is bouncy, Erin is bright.
- Since Erin is bright, Erin is friendly.
- Since Erin is friendly, Erin is purple.
Now, let's evaluate each option:
A) Erin is curious. - This cannot be inferred because the rules do not connect being smart or sleepy to being curious.
B) Erin is tasty. - This cannot be inferred because there is no information linking being bouncy, bright, or friendly to being tasty.
C) Erin is purple. - This can be inferred because Erin is friendly, and friendly people are purple.
D) Erin is clean. - This cannot be inferred because there is no information linking being bouncy, bright, or friendly to being clean.
Answer: C

Figure 6: Prompt template and model (Qwen) output (Entailment).

Prompt: input and output (Independence)

System: You are an assistant tasked with solving multiple-choice questions that require logical reasoning over the supplied knowledge diagrams. Use only the information explicitly given—do not rely on outside or commonsense knowledge. Read the question and given information, think step-by-step and answer the question. At the end of your answer, answer precisely in the format 'Answer: X' where X is the chosen letter A / B / C / D.

Input:



Dan is sleepy.

(Audio information: *Erin is spiky.*)

Rules are as follows: Blue person is tasty. Red person is clean. If a person is smart or sleepy, then the person is curious. Friendly person is purple.
Which of the following options can be inferred based on the given facts and rules? A) Erin is curious. B) Erin is purple. C) Erin is tasty. D) Erin is clean.

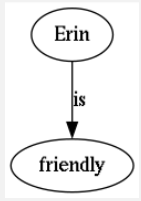
Output: To determine the correct answer, let's analyze the given information step by step:
1. Erin is friendly. 2. Friendly person is purple. 3. Blue person is tasty.
From the rules provided: - If a person is smart or sleepy, then the person is curious. - Red person is clean.
Since Erin is friendly and friendly people are purple, we can infer that Erin is purple.
Now, let's evaluate the options:
A) Erin is curious. - This cannot be inferred from the given information.
B) Erin is purple. - This can be inferred from the given information.
C) Erin is tasty. - This cannot be inferred from the given information.
D) Erin is clean. - This cannot be inferred from the given information.
Answer: B

Figure 7: Prompt template and model (Qwen) output (Independence).

Prompt: input and output (Contradictory)

System: You are an assistant tasked with solving multiple-choice questions that require logical reasoning over the supplied knowledge diagrams. Use only the information explicitly given—do not rely on outside or commonsense knowledge. Read the question and given information, think step-by-step and answer the question. At the end of your answer, answer precisely in the format 'Answer: X' where X is the chosen letter A / B / C / D.

Input:



Erin is red

 (Audio information: *Erin is blue*)

Rules are as follows: Blue person is tasty. Red person is clean. If a person is smart or sleepy, then the person is curious. Friendly person is purple.
Which of the following options can be inferred based on the given facts and rules? A) Erin is curious. B) Erin is tasty. C) Erin is purple. D) Erin is clean.

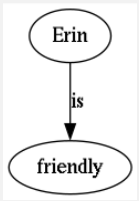
Output:
Let's analyze the information step by step:
1. Erin is blue.
2. Blue people are tasty.
3. Erin is friendly.
4. Friendly people are purple.
From the rules, we can infer the following:
- Since Erin is blue, she is tasty.
- Since Erin is friendly, she is purple.
Now, let's evaluate each option:
A) Erin is curious. - This cannot be inferred because the rules do not connect being blue or friendly to being curious.
B) Erin is tasty. - This can be inferred because blue people are tasty.
C) Erin is purple. - This can be inferred because friendly people are purple.
D) Erin is clean. - This cannot be inferred because the rules do not connect being blue or friendly to being clean.
Based on the given facts and rules, the options that can be inferred are B and C. However, since the question asks for a single answer, we need to choose the one that is directly supported by the rules without additional assumptions.
Answer: B

Figure 8: Prompt template and model (Qwen) output (Contradictory).

Prompt: input and output (Complementary)

System: You are an assistant tasked with solving multiple-choice questions that require logical reasoning over the supplied knowledge diagrams. Use only the information explicitly given—do not rely on outside or commonsense knowledge. Read the question and given information, think step-by-step and answer the question. At the end of your answer, answer precisely in the format 'Answer: X' where X is the chosen letter A / B / C / D.

Input:



Erin is purple

(Audio information: *Erin is red*)

If a person is purple and red and not friendly, then the person is soft. If a person is friendly and purple and not red, then the person is big. If a person is friendly and red and not purple, then the person is scary. If a person is friendly and purple and red, then the person is clean.
Which of the following options can be inferred based on the given facts and rules? A) Erin is soft. B) Erin is scary. C) Erin is clean. D) Erin is big.

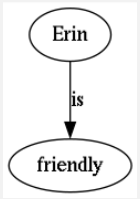
Output:
To determine which option can be inferred based on the given facts and rules, let's analyze the information step by step:
1. Erin is purple.
2. Erin is friendly.
3. Erin is red.
Now, let's apply the rules to Erin:
- Rule 1: If a person is purple and red and not friendly, then the person is soft. - Erin is purple and red and friendly, so this rule does not apply to Erin.
- Rule 2: If a person is friendly and purple and not red, then the person is big. - Erin is friendly, purple, and red, so this rule does not apply to Erin.
- Rule 3: If a person is friendly and red and not purple, then the person is scary. - Erin is friendly, red, and purple, so this rule does not apply to Erin.
- Rule 4: If a person is friendly and purple and red, then the person is clean. - Erin is friendly, purple, and red, so this rule applies to Erin.
Based on the analysis, the only rule that applies to Erin is Rule 4, which states that if a person is friendly and purple and red, then the person is clean.
Answer: C

Figure 9: Prompt template and model (Qwen) output (Complementary).

Prompt: input and output (Recognition)

System: You are an assistant tasked with solving multiple-choice questions about knowledge diagrams. Use only the information explicitly given—do not rely on outside or commonsense knowledge. The facts are given in image, audio and text. Read the question and given information, and directly answer the question in the following format: 'Answer: X' where X is the chosen letter A / B / C / D.

Input:



Dan is sleepy.

(Audio information: *Erin is spiky.*)

Question: Which fact is mentioned in the given information in image, audio or text?
Options: A) Erin is sticky., B) Erin is friendly., C) Erin is scary., D) Erin is green.

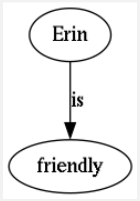
Output:
B

Figure 10: Prompt template and model (Qwen) output for the fact recognition.

Prompt: input and output (Two-Step Reasoning: Step 1)

System Prompt 1: You are an assistant tasked with solving multiple-choice questions about knowledge diagrams. Use only the information explicitly given—do not rely on outside or commonsense knowledge. The facts are given in image, audio and text. Read the question and given information, and directly answer the question.
System Prompt 2: You are an assistant tasked with solving multiple-choice questions that require logical reasoning over the supplied knowledge diagrams. Use only the information explicitly given—do not rely on outside or commonsense knowledge. Read the question and given information, think step-by-step and answer the question. At the end of your answer, answer precisely in the format 'Answer: X' where X is the chosen letter A / B / C / D.

Input 1:



Dan is sleepy.

(Audio information: *Erin is spiky.*)

Question: Check the given information and list all the facts in the given image, audio and text, respectively.

Output 1:
Facts from the image: - Erin is spiky.
Facts from the audio: - Dan is sleepy.
Facts from the text: - Erin is friendly.

Prompt: input and output (Two-Step Reasoning: Step 2)

Input 2 (Including Input 1 + Output 1):
Rules are as follows: Blue person is tasty. Red person is clean. If a person is smart or sleepy, then the person is curious. Friendly person is purple.
Which of the following options can be inferred based on the given facts and rules? A) Erin is curious. B) Erin is purple. C) Erin is tasty. D) Erin is clean.

Output 2:
Let's analyze the information step by step:
1. Erin is friendly.
2. Erin is spiky.
3. Dan is sleepy.
From the rules provided:
- If a person is smart or sleepy, then the person is curious. - Friendly person is purple.
- Blue person is tasty. - Red person is clean.
Since Erin is friendly, according to the rules, Erin must be purple.
Now, let's evaluate the options:
A) Erin is curious. - We don't have enough information to infer this from the given facts and rules.
B) Erin is purple. - This is directly inferred from the rule that a friendly person is purple.
C) Erin is tasty. - We don't have enough information to infer this from the given facts and rules.
D) Erin is clean. - We don't have enough information to infer this from the given facts and rules.
Answer: B

Figure 11: Prompt template and model (Qwen) output for the two step reasoning.

C SUPPLEMENTARY RESULTS

C.1 MODALITY INTERACTION: EQUIVALENCE

Setup. We now evaluate the *Equivalence* interaction, where the same decisive fact is redundantly placed in all three modalities (vision, audio, text). The unimodal conditions serve as baselines, where the only fact is presented in one modality. This setting examines whether redundant presentation across modalities can reinforce fact recognition and improve reasoning robustness.

Table 5: **Performance on the *Equivalence* interaction.** Each modality redundantly contains the same decisive fact. Results show that adding equivalent facts from other modalities provides no consistent improvement, suggesting that unimodal fact recognition is not the reasoning bottleneck.

Accuracy (%)	Unimodal			Multimodal $(\equiv)_{\Delta\mathbf{V},\Delta\mathbf{A},\Delta\mathbf{T}}$
	V	A	T	
Baichuan	79.4	75.0	95.9	$84.8_{5.4\uparrow, 9.8\uparrow, 11.1\downarrow}$
Qwen	96.3	94.4	98.0	$98.9_{2.6\uparrow, 4.5\uparrow, 0.9\uparrow}$
MiniCPM	89.4	89.6	95.0	$94.8_{5.4\uparrow, 5.2\uparrow, 0.2\downarrow}$
Phi4	58.8	60.2	96.6	$84.1_{25.3\uparrow, 23.9\uparrow, 12.5\downarrow}$
Average	91.0	79.8	96.4	$90.7_{9.7\uparrow, 10.9\uparrow, 5.7\downarrow}$

Results. As shown in Tab. 5, adding equivalent facts in vision and audio brings little to no gain over the text-only setting, which already achieves near-perfect accuracy. In some cases (e.g., Baichuan and Phi4), the multimodal setting even leads to a significant performance drop. This suggests that additional modalities do not enhance reasoning and may introduce unnecessary interference. Since models already recognize text-based facts effectively, improving recognition robustness does not translate into better multimodal reasoning. The bottleneck appears to lie elsewhere, specifically, in how the recognized facts are composed during reasoning.

Takeaway. Adding redundant evidence across modalities fails to improve reasoning and can even hurt performance. This suggests that text-based fact recognition is not the limiting factor. Instead, cross-modal fusion and composition seem to be the main challenges in multimodal reasoning. *Since simply reinforcing facts does not help, we next ask: can additional modalities help by introducing new and independent reasoning paths?* We investigate this in the following setting.

C.2 MODALITY INTERACTION: ALTERNATIVE

Setup. In this setting, each modality contains a unique antecedent that is individually sufficient to trigger the same rule and yield the correct conclusion. Thus, the model can arrive at the correct answer by reasoning over any one of the modalities. This evaluates whether models can flexibly leverage semantically diverse cues when multiple independent reasoning paths are available.

Table 6: **Performance on the Alternative interaction.** Each modality provides an independent reasoning path, and the correct answer can be inferred from any one of them. Multimodal accuracy improves slightly over the text-only baseline, showing that models can benefit from semantically diverse cues across modalities. This suggests that introducing complementary reasoning paths can help mitigate modality-specific limitations.

Accuracy (%)	Unimodal			Multimodal $(\vee)_{\Delta\mathbf{V},\Delta\mathbf{A},\Delta\mathbf{T}}$
	V	A	T	
Baichuan	78.0	79.8	97.3	$97.6_{19.6\uparrow, 17.8\uparrow, 0.3\uparrow}$
Qwen	96.3	93.9	97.4	$100.0_{3.7\uparrow, 6.1\uparrow, 2.6\uparrow}$
MiniCPM	92.0	91.1	96.2	$99.1_{7.1\uparrow, 8.0\uparrow, 2.9\uparrow}$
Phi4	77.6	71.6	96.9	$97.9_{20.3\uparrow, 26.3\uparrow, 1.0\uparrow}$
Average	86.0	83.9	97.0	$98.7_{12.7\uparrow, 14.8\uparrow, 1.7\uparrow}$

Results. As shown in Tab. 6, models achieve strong performance across unimodal settings, particularly in the text-only case. Notably, when all three sufficient premises are presented across modalities, performance consistently improves across models. For example, Qwen improves from 97.4% (text-only) to 100.0%, and MiniCPM improves from 96.2% to 99.1%. Although the gain is modest compared to the text baseline, the upward trend suggests that models can effectively integrate multiple alternative cues to reinforce the reasoning process, even when distributed across modalities.

Takeaway. These findings suggest that unlike redundant (Equivalence) setups, providing alternative reasoning paths across modalities can support more robust reasoning. Multimodal information helps when it offers diverse routes to the same conclusion, rather than simply reiterating or fragmenting the required information. Since introducing alternative single-step reasoning cues across modalities leads to measurable gains, it raises a natural follow-up question: can models also benefit from more indirect, multi-hop cues, such as entailment chains spread across modalities? We explore this next.

C.3 Modality Interaction: Entailment

Setup. In the Entailment interaction, the model must perform multi-hop reasoning through a chain of rules. Specifically, three facts form a reasoning chain: $A \rightarrow B \rightarrow C$, where the final answer is entailed by C. While the decisive fact C is always placed in one fixed modality (vision, audio, or text), the earlier support facts A and B are distributed across the other two modalities. This setting evaluates whether the model can integrate indirect, cross-modal evidence to support the final-step reasoning.

Table 7: **Performance on the Entailment interaction.** Each modality carries a fact needed for multi-hop reasoning, with the final-step premise fixed in one modality (V/A/T). Models consistently perform worse in multimodal setups compared to their unimodal baselines. This suggests that spreading multi-hop reasoning steps across modalities introduces substantial integration errors, regardless of where the final step is placed.

Accuracy (%)	Unimodal			Multimodal (→): Final-Step Fact		
	V	A	T	$V_{\Delta V}$	$A_{\Delta A}$	$T_{\Delta T}$
Baichuan	81.5	82.0	94.3	$79.5_{2.0\downarrow}$	$75.6_{6.4\downarrow}$	$80.7_{13.6\downarrow}$
Qwen	94.1	94.8	96.7	$78.4_{15.7\downarrow}$	$86.6_{8.2\downarrow}$	$83.9_{12.8\downarrow}$
MiniCPM	93.2	92.9	95.2	$81.8_{11.4\downarrow}$	$80.0_{12.9\downarrow}$	$88.4_{6.8\downarrow}$
Phi4	75.2	70.0	97.7	$73.0_{2.2\downarrow}$	$69.3_{0.7\downarrow}$	$79.7_{18.0\downarrow}$
Average	86.0	84.9	96.0	$78.2_{7.8\downarrow}$	$77.9_{7.1\downarrow}$	$83.2_{12.8\downarrow}$

Results. As shown in Tab. 7, all models experience significant performance drops in the multimodal condition compared to their unimodal counterparts, regardless of which modality carries the final-step fact. For instance, when the final-step fact is in text, we observe up to 12.8% drop compared to the text-only baseline. This pattern holds consistently across models and configurations. These results suggest that the cross-modal composition of reasoning chains introduces substantial integration errors, even when the decisive premise remains in a strong reasoning modality.

Takeaway. Unlike alternative information, which offers parallel reasoning paths, entailment information provides indirect, chained support, which proves to be not only unhelpful but actively harmful to reasoning performance. This highlights that current MLLMs struggle with cross-modal multi-hop reasoning, even when all required facts are present. We have now evaluated three types of auxiliary information: equivalence, alternative, and entailment, and found that only alternative reasoning paths offer modest improvements. In contrast, redundant or indirect information often introduces confusion and leads to performance degradation.

C.4 Modality Interaction: Independence

Tab. 8 presents detailed results under the *Independence* interaction, where each instance contains a single decisive fact placed in one modality (vision, audio, or text), while the other modalities contain only irrelevant distractors. We evaluate both unimodal reasoning (all facts are given in one modality) and multimodal reasoning (the decisive fact is mixed with distractors across modalities).

Table 8: **Performance on the *Independence* interaction with cross-modal distractors.** Each instance contains one decisive fact placed in a specific modality (V: vision, A: audio, T: text), while the other modalities contain only noisy facts. We report accuracy (%) for each unimodal condition as well as multimodal reasoning when decisive facts are distributed.

Accuracy (%)	**Unimodal**			**Multimodal (Decisive Facts)**			
	V	**A**	**T**	**V**	**A**	**T**	**Random**
Baichuan	60.2	72.0	94.8	74.3	53.5	74.9	67.6
Qwen	73.3	94.3	95.5	50.8	90.8	84.1	75.2
MiniCPM	77.6	83.7	91.2	66.8	78.2	91.0	78.7
Phi4	49.9	48.9	96.3	58.0	50.4	70.7	59.7
Average	65.3	74.7	94.5	62.5	68.2	80.2	70.3

Across all models, text consistently yields the highest unimodal accuracy (average 94.45%), while vision and audio vary significantly in performance. Interestingly, when the decisive fact remains in a fixed modality but is surrounded by irrelevant facts from other modalities, we observe substantial performance drops for vision and audio (e.g., Qwen drops from 73.3% to 50.8% when adding distractors to vision). The "Random" column averages over all cross-modal settings with randomly selected decisive modalities, showing that even a single irrelevant modality can degrade reasoning.

LLM USAGE

We used ChatGPT as a general-purpose assistant in preparing this paper. In particular, LLMs were employed for grammar refinement, clarity improvements, LaTeX formatting, and debugging minor code snippets. They were not involved in research ideation, experimental design, or the development of theoretical contributions.