Seeing vs. Believing: Evaluating the Language Bias of Open-Source MLLMs in Counter-Intuitive Scenes

Chen Ling[1*], **Tongwei Zhang**[2*], **Hanqian Li**[3*], **Nai Ding**[1†]
[1] College of Biomedical Engineering & Instrument Science, Zhejiang University
[2] Beijing University of Posts and Telecommunications
[3] Hong Kong University of Science and Technology (Guangzhou)

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated remarkable performance in mainstream visual understanding tasks, but their ability to process action scenes that contradict everyday common sense remains undertested. To address this gap, we introduce CAIT, a benchmark comprising 400 high-fidelity synthetic scenes focused on counter-intuitive visual actions, such as "a rabbit is chasing a tiger", where visual evidence explicitly contradicts common-sense expectations. We evaluate human, leading proprietary models (e.g., Claude and Gemini), and 14 representative open-source MLLMs. Humans achieve near-perfect performance (around 0.95 accuracy) and proprietary models demonstrate robust understanding (achieving up to 0.88 accuracy), standard open-source instruction-tuned models perform at the chance level. Further analysis demonstrates that this failure is driven by a strong language prior: rather than trusting the visual input, they automatically override the anomalous visual signals with statistically common text descriptions. Although introducing Chain-of-Thought reasoning mechanisms can improve accuracy, it significantly slows down the response and generates a new failure mode: models overthink the scenario and refuse to accept the actual visual content simply because it violates real-world physical laws. Finally, we demonstrate that targeted fine-tuning and structured prompting can effectively mitigate this reliance on language priors, enabling open-source models to accurately ground their reasoning in actual visual evidence. Our dataset is publicly available at https://huggingface.co/datasets/LukeLing/CAIT.

1 Introduction

Multimodal Large Language Models (MLLMs) have made remarkable breakthrough in multimodal understanding tasks. They have shown excellent performance in tasks like Question Answering (QA)(Fan et al., 2026; Min et al., 2026), Visual Question Answering (VQA) (Goyal et al., 2017; Xun et al., 2026; Tao et al., 2025; Shi et al., 2026; Yu et al., 2025), Image Captioning, and Visual Common Sense Reasoning (Zellers et al., 2019). Despite their impressive generalization abilities across large-scale standard datasets, their reasoning capabilities in counter-intuitive action scenes are undertested.

Figure 1: Example of an counter-intuitive action scene. Human responses comes from one of the annotators and Model Response is generated by GPT-4o-mini.

The disparity between human and model performance in these scenarios can be understood

*These authors contributed equally to this work.
†Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 1695–1712
October 24-29, 2026 ©2026 Association for Computational Linguistics

through **Dual Process Theory** (Krämer, 2014), which posits two modes of thought:

- **System 1** (Intuition): A fast, automatic, and "blink-level" judgment mode.
- **System 2** (Deliberation): A slow, logical, and computationally expensive reasoning mode

For humans, identifying a counter-intuitive action—such as "a bear cub bottlefeeds a zookeeper" in Figure1—is primarily a System 1 task. Humans possess a "neural reflex" for interaction anomalies, allowing them to identify logical or physical violations almost instantaneously. This process relies on a deep, implicit accumulation of social and cultural common sense(Huang et al., 2025; Hu et al., 2025).

In contrast, while current MLLMs are increasingly designed with analogous architectures (Hagendorff et al., 2023), they fail to replicate human efficiency. Even with System 2 thinking, they still often rely on language patterns for superficial associations rather than truly understanding the semantic relationships or action logic between objects in image (Niu et al., 2021). Numerous studies point out the vulnerabilities of MLLMs in semantic reversal scenarios (Thrush et al., 2022) and common sense conflict scenarios.(Bitton-Guetta et al., 2023) However, they mostly focus on object-level counterfactuals or text semantic conflicts, with little attention given to deep visual reasoning at the action level.

To address this, we introduce **CAIT** (**C**ounter-Intuitive **A**ction **I**mage-**T**ext dataset), a multimodal VQA benchmark specifically designed to evaluate whether MLLMs can overcome language priors to comprehend counter-intuitive action logic. Utilizing image synthesis pipeline, we construct scenarios that challenge models to move beyond pattern matching and perform genuine object-relationship reasoning. Our main contributions are as follows:

- We present a novel VQA benchmark that systematically evaluates MLLMs' understanding of counter-intuitive visual actions and object-interaction relationships.
- We propose a semi-automated pipeline using state-of-the-art text-to-image model to generate high-fidelity, semantically consistent images with fine-grained VQA annotations.
- We provide an in-depth analysis of why current open-source MLLMs fail on CAIT and offer preliminary strategies to improve model performance in these challenging scenarios.

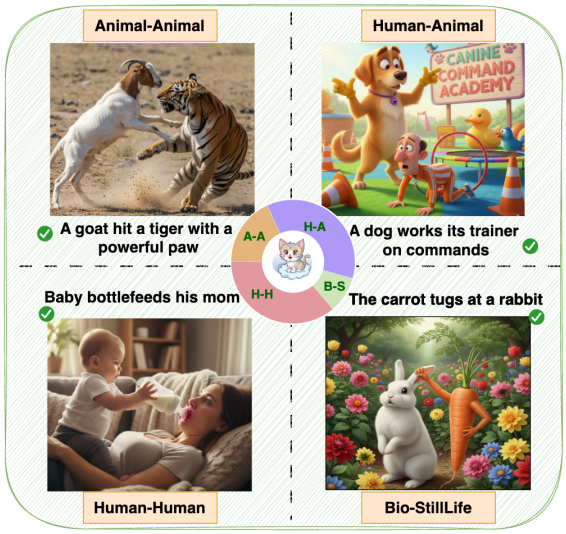

Figure 2: The four main classifications and corresponding examples of **CAIT**.

2 Related Work

Evolution of MLLMs The field of multimodal reasoning has shifted significantly toward integrating Large Language Models (LLMs) with visual encoders to form MLLMs. Early breakthrough such as Flamingo (Alayrac et al., 2022), demonstrated remarkable few-shot (Brown et al., 2020) capabilities in cross-modal reasoning through gated cross-attention mechanisms. Building on this, the current research landscape is dominated by models like LLaVA (Liu et al., 2023), Qwen-VL (Bai et al., 2023), and InstructBLIP (Dai et al., 2023), which utilize visual feature projection and large-scale instruction tuning to align visual signals with the semantic space of LLMs. More recently, a new generation of "Thinking" MLLMs has emerged, such as Qwen-VL-Thinking series (Bai et al., 2025a). These models aim to simulate deliberative cognition by incorporating internal monologues or Chain-of-Thought (CoT) (Wei et al., 2023) processes before generating final outputs, specifically targeting complex logical reasoning tasks.

Counterfactual Reasoning Benchmarks Counterfactual reasoning are essential for verifying whether MLLMs genuinely understand visual content or merely exploit statistical biases. To address the reliance on language priors, Winoground (Thrush et al., 2022) assesses visual-text alignment under semantic inversion by constructing

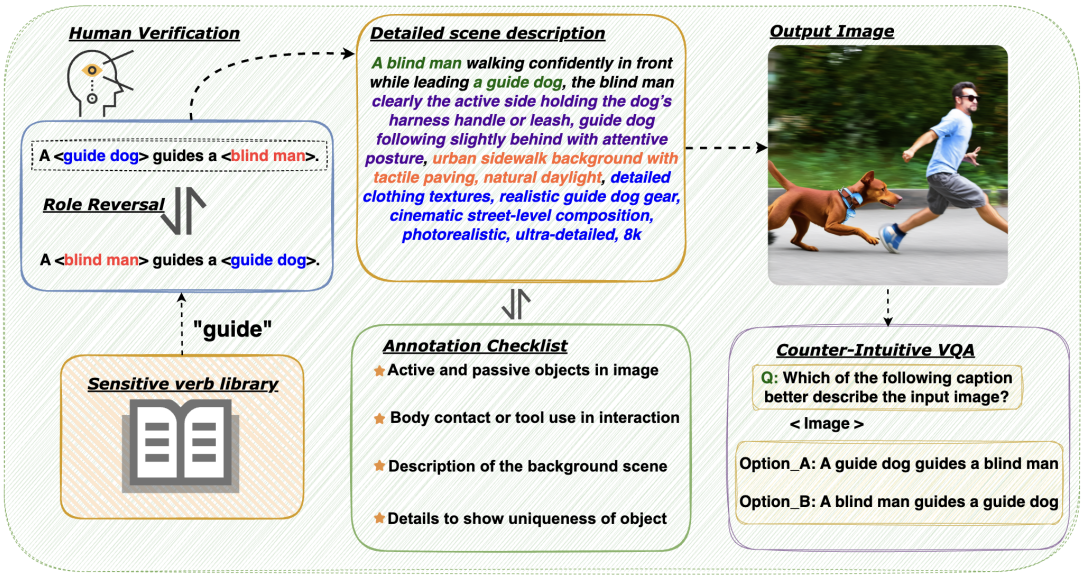

Figure 3: The construction framework of an counter-intuitive action image-text pair.

matching tasks with semantically swapped text descriptions. Contrastive evaluation methods such as VQA-CP (Agrawal et al., 2018) construct uncommon distribution by disrupting the answer distribution consistency between training and test sets. However, these methods do not strictly involve genuine uncommon image scenes. More recently, COCO-Counterfactual (Le et al., 2023) tests model sensitivity to visual changes by modifying objects or attributes within images, while CFMM (Li et al., 2025) evaluates reasoning capabilities through systematic counterfactual rewriting of question text, but they are predominantly limited to single-modality counterfactual construction.

Commonsense Conflict and Interaction Logic While benchmarks like WHOOPS! (Bitton-Guetta et al., 2023) introduce visual scenes that violate general commonsense, they often require complex, knowledge-intensive inference. Most related to our work is BLINK (Fu et al., 2024), which evaluates "blink-level" visual reasoning tasks. However, BLINK primarily targets core visual tasks such as spatial relations and counting. Our work fills a distinct gap by focusing specifically on interaction logic—the fundamental physical and social relationships between objects. Unlike WHOOPS!, which may rely on external cultural knowledge, CAIT emphasizes intuitive judgments derived from the agent-action-patient structure within a static frame. This focus on counter-intuitive actions allows for evaluation of how models balance visual signals against entrenched language priors.

3 Methodology

3.1 Overview

To systematically evaluate MLLM reasoning, we carefully construct CAIT through a generation pipeline that pairs high-fidelity images with rigorous semantic annotations.

Diverse Visual Action Scenes CAIT consists of 400 distinct scenarios organized into four taxonomic categories and six characteristic variations. Distribution of the scenes is detailed in Table 1.

Instant Thinking Mode The task is formulated to evaluate "neural reflex" perception, prioritizing immediate visual recognition over multi-step causal analysis. Unlike standard VQA, CAIT requires models to determine commonsense alignment directly from visual cues, bypassing the linguistic parsing of hypothetical conditions.

Blink-Level Tasks for Human As illustrated in Figure 1, the images in CAIT are designed to achieve "blink-level" difficulty. This establishes a human baseline where interaction anomalies are immediately perceptible, typically requiring few seconds for accurate identification.

High-Quality Synthetic Data We leverage Gemini-2.5-flash-image (Comanici et al., 2025) to generate images, which minimizes data contamination risks associated with web-crawled datasets. Each sample undergoes human verification to maintain high semantic fidelity and visual clarity.

Category	Ratio	Key Characteristics
Human-Human	22.00%	Inversion of institutional power
	14.25%	Reversal of caretaking duties across generations
Human-Animal	22.25%	Animals replacing humans as dominant figures
	14.00%	Animals providing domestic care to humans
Animal-Animal	18.00%	Transposition of traditional predator and prey roles
Bio-StillLife	9.50%	Inanimate objects gaining autonomy

Table 1: Taxonomy and distribution of counter-intuitive scenes in the CAIT dataset.

3.2 Data Processing

The data processing pipeline is divided into three phases: Verb Library Filtering, Counter-intuitive Image-Text Generation, and VQA Dataset Design.

Verb Library Filtering Using VerbNet (Kipper Schuler et al., 2009) as a foundation, we filter verbs based on four core principles, see detailed examples in Appendix A.1: **1) Semantic Consistency:** Verbs must explicitly describe the action relationship between objects. **2) Semantic Non-interchangeability:** The verb must follow "Agent-Action-Patient" structure, meaning that swapping the subject and object results in a fundamental change in the scenario's meaning or logic. **3) Static Recognizability:** Guided by existing visual action definitions(Ronchi and Perona, 2015), actions must be clearly discernible in a single static frame without temporal or contextual cues. **4) Commonality:** Verbs should be high-frequency and general-purpose, strictly avoiding domain-specific jargon.

Following these principles, we construct the ***Sensitive-Action Verb Library***, consisting of 53 categories and a total of 318 verbs. Table 2 presents the verb classification and corresponding judgment.

Class	Verb	Judgement Basis
6	40	direction of agent
19	106	active, passive, body contac
12	80	active, passive, tool using
12	80	active, passive, body part
4	21	position of agent

Table 2: Classification of *Sensitive-Action Verb Library*

Counter-intuitive Image-Text Generation We begin by selecting verbs from the sensitive-action library to construct counter-intuitive texts. Specifically, standard sentences with a Agent-Verb-Patient structure typically convey intuitive and commonsense meanings commonly found in daily language. However, the selected verbs are semantically non-interchangeable. That is, swapping the agent and patient significantly alters the sentence's meaning, resulting in a counter-intuitive scenario.

To efficiently collect a larger scale of such sentences and their corresponding images, we designed a data synthesis framework as illustrated in Figure 3. **1) Sentence Generation**: We prompt GPT-4o-mini to generate semantic scene pairs derived from specific verbs (see Table 19). For each verb, the model produces both a normal, commonsense sentence and a corresponding counter-intuitive version by swapping the subject and object, creating a logical but extremely uncommon scenario. **2) Human Verification**: We implement a human-in-the-loop verification process using a custom GUI (see Table 20). Human annotators review the raw generated pairs to filter out nonsensical inversions and configure essential parameters for each scene, specifically determining whether to anthropomorphize the subject and selecting the visual style (realistic or cartoon). **3) Scene Description**: We then prompt GPT-4o-mini to generate highly detailed and visually grounded descriptions (see Table 21). These descriptions are tailored to clearly convey the subject–object role reversal while adhering to the specified style and anthropomorphism settings. **4) Image Generation and Filtering**: Finally, we employ Gemini-2.5-flash-image to generate images based on the refined scene descriptions (see Table 22). Human annotators then rigorously review and filter the output to ensure each image-text pair clearly conveys the counter-intuitive concept while removing ambiguous or distracting elements.

VQA Dataset Design The resulting CAIT benchmark comprises 400 images paired with specific questions for MLLM evaluation. As illustrated in Figure3, each image is paired with a single ques-

tion where the question stem remains constant, and only the answer choices vary. The two primary options consist of the original commonsense text and the counter-intuitive text where the subject and object are swapped.

3.3 Human Annotation for VQA

We invited five annotators to perform manual annotations, establishing a high-quality comparison baseline for existing MLLMs. The annotation process records the time interval between the initial presentation of a question and the final submission as the reasoning time. Experimental results demonstrate that even if participant over 50 years old can achieve an average reasoning time of around 5 seconds. This consistently validates that the CAIT dataset achieves its "blink level" design goal, allowing humans to make near-instantaneous judgments. Detailed information is provided in Appendix A.2.

4 Experiment

4.1 Model Selection

To comprehensively evaluate the reasoning capabilities of MLLMs in CAIT scenarios, we selected leading proprietary models (Claude-4.6-opus and Gemini-2.5-pro) and 14 representative open-source models, categorized by their parameter scale and reasoning paradigms.

Instruct MLLMs Instruct-based models are fine-tuned to follow human instructions directly, representing the System 1 thinking mode. Our selection includes 3B, 7B and 72B version of Qwen2.5-VL-Instruct (Bai et al., 2025b), 4B, 8B and 235B version of Qwen3-VL-Instruct (Bai et al., 2025a), 4B and 12B version of Gemma-3-it (Team et al., 2025a), LLaVA-NeXT-7B (Li et al., 2024), and Llama-3.2-Vision-11B(Grattafiori et al., 2024).

Thinking MLLMs Thinking models are designed to simulate deliberative cognition by incorporating an explicit internal reasoning process before generating the final output, representing the System 2 thinking mode. We evaluate 4B and 8B version of Qwen3-VL-Thinking and InternVL-3.5, GLM-4.1V-9B-Thinking (Team et al., 2026) and Kimi-VL-A3B-Thinking(Team et al., 2025b).

4.2 Implementation Details

We set up two tasks in order to evaluate not only the primary VQA performance, but also the semantic understanding consistency of MLLMs. To establish an ideal performance upper bound for the evaluation, we set reference baselines for both categories. In the Counter-intuitive VQA task, we utilize the average performance of human annotators (see 3.3) as the human baseline to measure the gap in models' intuitive judgment. For the Gemini-re-eval task, we use the **GT Prompt** as input and the judgment results of Gemini-2.5-pro as the ground-truth baseline to verify the reasoning upper bound at the semantic description level.

Counter-intuitive VQA This task evaluates the model's capacity when confronted with visual actions that violate common sense. To provide a comprehensive assessment, we utilize Accuracy (Acc) and $F1$-score as the primary evaluation metrics:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i) \tag{1}$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

where N denotes the total number of test samples, $\hat{y}_i$ and y_i represent the predicted and ground-truth options for the i-th instance.

To deeply analyze computational overhead and logical depth, we precisely record the inference time (t_{infer}) for each question. To eliminate measurement errors caused by hardware variations, all open-source models are deployed on a single NVIDIA A100 (80GB) GPU using the vLLM framework (Kwon et al., 2023) for online inference services. In the inference settings, we uniformly adopt a greedy decoding strategy with a temperature of $T = 0.0$ to ensure output consistency. This process can be directly expressed as:

$$\mathcal{A}_{direct} = f_{\text{MLLM}}(I, Q) \tag{3}$$

where I is the original image, Q is the textual question, and $\mathcal{A}_{direct}$ is the directly output A/B option.

For Instruct MLLMs, in addition to the standard "instant answer", we introduce a CoT variant experiment to simulate the System 2 reasoning mode. In this setting, we guide the model to generate a reasoning chain before providing the final option. This aims to explore whether the failure of instruction-tuned models in these scenarios stems from initial misdirection at the visual perception layer or a lack of subsequent logical integration capabilities.

LLM-re-eval To further deconstruct whether the model has truly captured the interaction semantics

Model	Param	Counter-intuitive VQA			LLM-re-Eval	
		Acc	F1	Time	Acc	F1
Upper Bound						
Human Annotators	–	0.948	0.948	3.740	–	–
GT Prompt	–	–	–	–	0.984	0.985
Leading proprietary MLLMs						
Gemini-2.5-pro	–	0.885	0.886	0.804	0.835	0.835
Claude-4.6-opus	–	0.863	0.863	0.817	0.788	0.788
GPT-5.4	–	0.805	0.804	0.832	0.778	0.778
Grok-4.3	–	0.893	0.892	0.819	0.788	0.789
Open-source Instruct MLLMs						
Qwen2.5-VL-Instruct	7B	0.420	0.347	0.940	0.660	0.660
Qwen3-VL-Instruct	8B	0.380	0.327	0.853	**0.762**	**0.762**
Gemma-3-it	12B	0.360	0.353	0.843	0.596	0.596
LLaMA3.2-Vision	11B	0.375	**0.372**	2.622	0.607	0.607
LLaVA-NeXT	7B	**0.475**	0.349	1.870	0.548	0.548
Qwen2.5-VL-Instruct	3B	0.373	0.300	0.924	0.584	0.585
Qwen3-VL-Instruct	4B	0.328	0.326	0.633	0.707	0.707
Gemma-3-it	4B	0.328	0.334	**0.819**	0.541	0.541
Open-source Thinking MLLMs						
Qwen3-VL-Thinking	8B	0.640	0.635	10.645	0.696	0.696
GLM-4.1V-Thinking	9B	0.630	0.647	**5.770**	**0.733**	**0.733**
Kimi-VL-A3B-Thinking	16B	**0.683**	**0.680**	7.396	0.689	0.689
InternVL-3.5	8B	0.545	0.545	106.512	0.655	0.655
Qwen3-VL-Thinking	4B	0.625	0.622	14.385	0.695	0.695
InternVL-3.5	4B	0.618	0.618	50.333	0.683	0.683

Table 3: Performance of MLLMs on CAIT. The LLM-re-Eval scores are averaged over the three proprietary judges (Gemini-2.5-pro, Claude-4.6-Opus and GPT-5.4); the per-judge results are reported in Table 15 of Appendix A.5. Values denoted in **bold** indicate the highest performance in each open-source model group.

between different objects, we design an indirect reasoning experiment. The main purpose is to decouple image perception from logical judgment. First, the model under test generates a text description based on the image, which is then input alongside the question into one of three proprietary judges Gemini-2.5-pro, Claude-4.6-Opus, or GPT-5.4 for secondary evaluation:

$$\mathcal{D} = f_{\text{MLLM}}(I), \quad \mathcal{A}_{eval} = g_{\text{Judge}}(\mathcal{D}, Q) \quad (4)$$

where g_{Judge} represents the discriminative model and $\mathcal{D}$ is the generated semantic description, we took the average accuracy and F1 of three Judges as the LLM-re-eval.

Since all counter-intuitive images in the CAIT are synthesized by advanced text-to-image model, we define the original text prompt used to synthesize the image as the *GT Prompt*. As the direct source of image generation, *GT Prompt* represents the most accurate and unambiguous textual description of the action logic within that image.

Experimental results indicate that when any of the three proprietary judges take the *GT Prompt* as input, average accuracy reaches approximately *98%*. Among the evaluated models, proprietary judges achieve the highest LLM-re-eval scores, while open-source Instruct and Thinking models exhibit substantially lower accuracies, highlighting challenges in capturing the semantic logic of counter-intuitive interactions. This result demonstrates three proprietary judges' superior semantic logic capabilities and verifies that the counter-intuitive VQA problem is logically resolvable provided the image description is sufficiently accurate. Consequently, a low score in LLM-re-eval di-

rectly reflects a model's failure to correctly identify counter-intuitive interaction relationships during the visual perception phase.

Model	Accuracy	F1 Score	Infer Time
∼10B Instruct MLLMs			
Qwen2.5-VL-7B-Instruct	0.483	0.472	3.060
Qwen3-VL-8B-Instruct	0.643	0.641	7.747
Gemma-3-12b-it	0.535	0.533	3.824
LLama3.2-vision-11B	0.498	0.509	7.713
LLaVA-NeXT-7B	0.458	0.452	3.537
∼3B Instruct MLLMs			
Qwen2.5-VL-3B-Instruct	0.485	0.468	2.296
Qwen3-VL-4B-Instruct	0.603	0.602	22.809
Gemma3-4B-it	0.478	0.477	1.666

Table 4: Results of Instruct models with CoT reasoning.

4.3 Main Result

Experimental results in Table 3 demonstrate that the performance of all evaluated open-source MLLMs is significantly lower than the upper bound when processing counter-intuitive action scenes. While proprietary models, show strong performance with accuracies up to 0.88 in VQA task, a substantial performance gap persists between open-source and proprietary models. This indicates that current open-source models still face substantial challenges in parsing CAIT scenarios.

The LLM-re-Eval scores in Table 3 are averaged over the three proprietary judges, whose per-judge breakdown is reported in Table 15 (Appendix A.5). The GT-Prompt upper bound reaches around 0.98 across all judges. Proprietary models achieve the highest accuracies, while open-source Instruct models range from 0.532 to 0.773 and Thinking models from 0.663 to 0.743. Results are consistent across judges, reflecting judge-independent evaluation.

Language bias of Instruct MLLMs Instruct MLLMs exhibit rapid inference speeds, reflecting System 1 fast-thinking traits, but the accuracy consistently perform at the chance level. This reveals that such models rely heavily on language priors. When confronted with conflicts between visual facts and common sense, they tend to overlook actual visual features in favor of common-sense biases, revealing a lack of genuine understanding regarding physical logic.

Gains and Costs of Thinking Experimental data clearly show that incorporating a thinking guidance mechanism effectively corrects erroneous judgments. For Instruct models in Table 4, performance generally improves upon the introduction of CoT reasoning. Native Thinking models similarly exhibit enhanced robustness. However, the average inference time per question is several times—or even orders of magnitude—higher than that of Instruct models. This suggests that such logical correction comes at a significant computational cost, highlighting the high inference premium of "slow thinking" in counter-intuitive reasoning.

Semantic Collapse in Gemini-re-Eval As demonstrated in 4.2, the logic of the problem itself is not inherently difficult for a powerful MLLM. However, when the input is replaced with captions generated by open-source models (except Qwen3-VL series), the performance of Gemini-2.5-pro drops significantly. This suggests that when describing images, models often automatically "correct" the observed counter-intuitive facts into erroneous common-sense descriptions during output, leading to a collapse of visual signals during semantic descriptions transformation.

5 Analysis

5.1 Performace across different domains

We further analyze the impact of various subjects on model accuracy, with average results summarized in Table 5 and details provided in Appendix A.6. The models generally outperform other domains within the Animal-Animal category, which may be attributed to the higher discriminability of visual features in predatory relationships at the lower levels of visual processing. Conversely, the Human-Human domain renders Instruct models particularly susceptible to being misled by entrenched social common-sense priors. Notably, a comparison between System 1 and System 2 models reveals that the Bio-StillLife domain yields the most substantial reasoning gain, demonstrating that explicit logical deduction in reasoning can effectively recover authentic visual signals from perceptual hallucinations in physically absurd scenarios.

5.2 Attempts to boost performance

Structured Prompt for System 2 Enhancement To elevate the reasoning ceiling of models in System 2 mode, we designed a Structured Prompt specifically tailored for scenarios in CAIT (see Appendix A.4). This strategy is intended to guide models in explicitly analyzing the interaction relationships between objects, arriving at conclusions

Model	H-H	H-A	A-A	B-S
Instruct MLLMs				
∼10B	0.323	0.433	0.447	0.389
∼3B	0.299	0.366	0.509	0.272
Thinking MLLMs				
∼10B	0.633	0.607	0.646	0.533
∼3B	0.631	0.559	0.590	0.619

Table 5: The average accuracy of open-source models across categories (A-A: Animal-Animal, B-S: Bio-StillLife, H-A: Human-Animal, H-H: Human-Human).

by questioning physical plausibility and deliberately evaluating option logic. Table 6 presents the average performance metrics, with results for individual models provided in Appendix A.7. The results indicate that nearly all models, with the exception of LLama3.2-vision-11B and Qwen2.5-VL-3B-Instruct, benefited from this strategy, with 10B-scale models demonstrating significant gains. Notably, Qwen3-vl series showed exceptional performance under structured prompting, with Qwen3-vl-thinking achieving a leap of nearly 14%. This suggests that reinforcing deliberative instructional guidance can effectively unlock the latent reasoning potential of MLLMs.

Model	**Accuary**	**F1 Score**
Instruct Models		
∼10B	0.554 ↑ 4.2%	0.546 ↑ 3.7%
∼3B	0.518 ↑ 1.3%	0.518 ↑ 3.0%
Thinking Models		
∼10B	0.688 ↑ 7.2%	0.687 ↑ 6.8%
∼3B	0.650 ↑ 3.8%	0.648 ↑ 3.6%

Table 6: The average accuary and F1 score of open-source models after structured prompt.

LoRA-SFT for System 1 Calibration To address the vulnerability of instruction-tuned models to language-prior interference in System 1 mode, we employed Supervised Fine-Tuning to directly calibrate the models' cognitive preferences. We partitioned the CAIT into a 60% training set and a 40% test set, utilizing LoRA (Hu et al., 2021) to fine-tune Qwen2.5-vl-Instruct model while freezing the vision encoder. This ensures that the models learn new semantic correspondences while maintaining robust visual feature extraction capabilities. Experimental results in Table7 show that both fine-tuned models achieved performance increments of over 15%. This proves that fine-tuning on domain-specific distributions can effectively mitigate reliance on common-sense language priors, enabling more accurate, near-instantaneous responses to counter-intuitive visual signals.

Model	**Accuracy**	**F1 score**
Qwen-2.5-vl-7B	0.425	0.322
Qwen-2.5-vl-7B-LoRA	0.575 ↑ 15.0%	0.473 ↑ 15.2%
Qwen-2.5-vl-3B	0.400	0.308
Qwen-2.5-vl-3B-LoRA	0.581 ↑ 18.1%	0.498 ↑ 18.9%

Table 7: Comparison of Qwen2.5-vl series during SFT.

5.3 Failure Analysis

We further investigate the failure cases from 5.2 and categorize them into four modes based on the source of hallucination: **1) Statistical Frequency Bias:** The model acknowledges the possibility of an event but rejects valid visual evidence because the scenario contradicts the statistical likelihoods observed in its training distribution. **2) Logical Impossibility:** Model employs a rigid rejection mechanism, asserting that visual content violates fundamental physical laws or logical constraints, often explicitly labeling the scenario as "absurd" or "impossible". **3) Semantic Binding Failure:** A pure vision-language alignment error where the model correctly identifies objects but misattributes semantic roles (e.g., Agent-Patient reversal) or spatial relation due to attention failures. **4) Alignment Failure:** A disconnect between reasoning and generation, where the model deduces the correct answer in its chain-of-thought but hallucinates an inverted final label due to the re-emergence of language priors.

Figure 4 shows the probability of each error. Also, we list the specific failure case in Appendix C. Instruct models are predominantly plagued by shallow *Statistical Frequency Bias* and *Spatial Binding Failures*, reflecting System 1's reliance on heuristics over fine-grained visual cues. While introducing reasoning capabilities significantly suppresses these errors, it introduces a paradox of overthinking; *Logical Impossibility* errors surge as models actively reject counter-intuitive visual truths, and *Alignment Failures* rise, indicating a disconnect where instinctive language priors override the correct deductive chain during final generation.

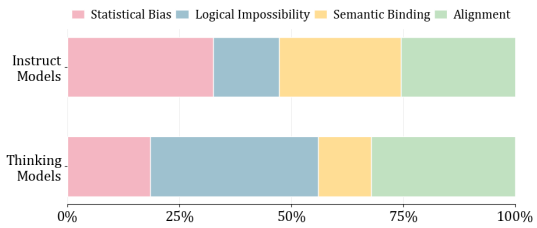

Figure 4: Distribution of error types of open-source MLLMs.

5.4 CAIT Dataset Quality Evaluation

To assess the quality of the CAIT dataset, we conducted two complementary evaluations: perceptual image quality using no-reference IQA metrics and semantic alignment using CLIP cosine similarity. This ensures model failures are due to reasoning deficits rather than image artifacts or misalignment.

Perceptual Quality (NR-IQA). We computed NIQE and BRISQUE scores for CAIT images to quantitatively evaluate their perceptual naturalness and presence of artifacts. The CAIT dataset achieves a mean NIQE score of 3.08 and a mean BRISQUE score of 8.12. In comparison, images from the Winoground dataset exhibit lower quality, with a mean NIQE of 4.33 and a BRISQUE of 32.88. These results indicate that CAIT images possess superior perceptual quality and fewer structural distortions than Winoground, ensuring that model errors are not a byproduct of poor image signal quality.

Dataset	NIQE ↓	BRISQUE ↓
CAIT	3.08	8.12
Winoground	4.33	32.88

Table 8: Perceptual quality (NR-IQA) for CAIT and Winoground. Lower score indicate better image quality.

Semantic Alignment (CLIP). Using CLIP ViT-B/32, we computed cosine similarity between image embeddings and captions. CAIT images achieve a mean correct-caption score of 0.295 and reversed-caption score of 0.295, yielding a near-zero semantic margin of -0.0007. For Winoground, the corresponding scores are 0.261, 0.258, and a margin of 0.0031. These results confirm that CAIT images are semantically aligned, and the near-zero margin emphasizes that models cannot rely solely on object recognition but must reason about interactions and compositional logic.

Dataset	Correct Caption	Reversed Caption
CAIT	0.295	0.295
Winoground	0.261	0.258

Table 9: Mean CLIP cosine similarity between images and captions for CAIT and Winoground datasets.

6 Conclusion

We introduce CAIT, a novel benchmark comprising 400 independent scenes designed to fill the current gap in counter-intuitive action reasoning. Our evaluation and analysis reveals a stark contrast: while human intuition and leading proprietary models perform exceptionally well, open-sourced models in System 1 mode perform at the chance level, indicating a strong reliance on language priors. Although System 2 reasoning provide notable performance gains, it slow down the responses and introduce new failure mode where models over-rationalize and explicitly reject visual truths. Furthermore, we demonstrate that targeted fine-tuning and structured prompting can effectively mitigate this language bias, providing insights for developing future multimodal models capable of better physical and interaction logic understanding.

Limitations

Monolingual Evaluation Currently, our benchmark is restricted to English. It is important to verify that our findings may generalize across various languages and cultural contexts. Our future work will expand the benchmark to multilingual scenarios such as Chinese.

Restriction to Static Imagery The scenes of our work are specifically within a single static frame. Extending to video-based counter-intuitive interactions would introduce dynamic challenges such as temporal and audio reasoning, providing a more comprehensive evaluation for MLLMs.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful suggestions and comments. This work was supported by the Brain Science and Brain-like Intelligence Technology — National Science and Technology Major Project 2021ZD0204100 (2021ZD0204105 to N.D.).

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. *Preprint*, arXiv:1712.00377.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. *Preprint*, arXiv:2303.07274.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Yuchen Fan, Chen Ling, Xin Zhong, Shuo Zhang, Heng Zhou, Yuchen Zhang, Mingyu Liang, Chengxing Xie, Ermo Hua, Zhizhou He, Cheng Huang, Gang Chen, Ning Ding, and Bowen Zhou. 2026. LFQA-e: Carefully benchmarking long-form QA evaluation. In *The Fourteenth International Conference on Learning Representations*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. *Preprint*, arXiv:2404.12390.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Vqa v2: A dataset for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Xuming Hu, Hanqian Li, Jungang Li, Yu Huang, Shuliang Liu, Qi Zheng, Junhao Chen, and Aiwei Liu. 2025. Videomark: A distortion-free robust watermarking framework for video diffusion models. *arXiv preprint arXiv:2504.16359*.

Sirui Huang, Hanqian Li, Yanggan Gu, Xuming Hu, Qing Li, and Guandong Xu. 2025. Hyperg: Hypergraph-enhanced llms for structured knowledge. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1218–1228.

Karin Kipper Schuler, Anna Korhonen, and Susan Brown. 2009. VerbNet overview, extensions, mappings and applications. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion*

Volume: Tutorial Abstracts, pages 13–14, Boulder, Colorado. Association for Computational Linguistics.

W. Krämer. 2014. Thinking, fast and slow. *Statistical Papers*, 55:915.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Tiep Le, Vasudev Lal, and Phillip Howard. 2023. Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. *Preprint*, arXiv:2309.14356.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *Preprint*, arXiv:2407.07895.

Yian Li, Wentao Tian, Yang Jiao, Jingjing Chen, Tianwen Qian, Bin Zhu, Na Zhao, and Yu-Gang Jiang. 2025. Look before you decide: Prompting active deduction of mllms for assumptive reasoning. *Preprint*, arXiv:2404.12966.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Shangchao Min, Shaonan Wang, Xinyu Gao, Hui Wang, Zhiling Jin, Chen Ling, and Nai Ding. 2026. Standardized assessment of llm english proficiency.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. *Preprint*, arXiv:2006.04315.

Matteo Ruggero Ronchi and Pietro Perona. 2015. Describing common human visual actions in images. *Preprint*, arXiv:1506.02203.

Yibo Shi, Jungang Li, Linghao Zhang, Zihao Dongfang, Biao Wu, Sicheng Tao, Yibo Yan, Chenxi Qin, Weiting Liu, Zhixin Lin, and 1 others. 2026. Androtmem: From interaction trajectories to anchored memory in long-horizon gui agents. *arXiv preprint arXiv:2603.18429*.

Sicheng Tao, Jungang Li, Yibo Yan, Junyan Zhang, Yubo Gao, Hanqian Li, ShuHang Xun, Yuxuan Fan, Hong Chen, Jianxiang He, and 1 others. 2025. Mosschatv: Reinforcement learning with process reasoning reward for video temporal reasoning. *arXiv preprint arXiv:2509.21113*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025a. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, and 76 others. 2025b. Kimi-vl technical report. *Preprint*, arXiv:2504.07491.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. *Preprint*, arXiv:2204.03162.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Shuhang Xun, Sicheng Tao, Jungang Li, Yibo Shi, Zhixin Lin, Zhanhui Zhu, Yibo Yan, Hanqian Li, Linghao Zhang, Shikang Wang, and 1 others. 2026. Rtv-bench: Benchmarking mllm continuous perception, understanding and reasoning through real-time video. *Advances in Neural Information Processing Systems*, 38.

Qianxi Yu, Chen Ling, Shitao Feng, Jiayi Hu, Yongtong Lu, and Nai Ding. 2025. Giics: A dataset for gaze-direction-based intention inference in complex visual scenes. In *2025 4th International Conference on Artificial Intelligence, Human-Computer Interaction and Robotics (AIHCIR)*, pages 1–9.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

A Supplementary Materials

A.1 Examples of Verb Library Filtering

To ensure the high quality and recognizability of the action scenes in CAIT, we followed four core principles during the filtering of the *Sensitive-Action Verb Library*. Table 10 provides specific examples of verbs that passed or failed these criteria.

A.2 Detailed Information of Annotators

Annotator Selection The selected annotators represent diverse age groups and educational backgrounds, with a minimum requirement of an elementary school education and a fundamental understanding of real-world common knowledge. By excluding professional annotation training, we ensure that the baseline reflects natural human intuition rather than specialized task-solving strategies.

Annotation Process and Compensation

- **Interface:** Annotations were conducted via a dedicated web interface that presents questions as individual, sequential items.
- **Task Flow:** Annotators received one question at a time and submitted their responses directly on the webpage to ensure a focused, rapid judgment process.
- **Ethics and Compensation:** To ensure high-quality data and ethical labor standards, each annotator was compensated at a rate of 7 dollars per hour, approximately 2.5 times the local minimum wage.

Annotation Statistics Table 11 shows detailed performance metrics for each annotator. The results demonstrate high accuracy and consistency across diverse groups, with an average Fleiss' κ of 0.880. The average response time of 3.740 seconds per item aligns with the "blink-level" judgment criteria, confirming that the task successfully captures rapid human intuition.

A.3 Prompts of MLLMs during Inference

Table 12 shows the original prompt for MLLMs. Table 13 shows the CoT prompt for Instruct Models

A.4 Structured prompts for System 2 thinking mode

Table 14 shows the structured prompt for System 2 mode.

A.5 Per-Judge Results of LLM-re-Eval

The LLM-re-Eval column reported in Table 3 is the average over the three proprietary judges. Table 15 reports the accuracy obtained with each individual judge, showing that the ranking of the evaluated models is largely independent of the judge being used.

A.6 Performance of Models across different domains

Table 16 shows the performance of specific models across different domains.

A.7 Performance of different models using structured prompts

Table 17 and Table 18 show the performance of different models using structured prompts.

B Prompt of Workflow Details

We present the core instructions for the workflow in Tables 19 to 22. This includes the task definition and the constraints provided to the models.

C Error Analysis

To better understand the limitations of current MLLMs, we conduct a qualitative error analysis. We random sample 200 error cases from the Structured Prompt experiments of both Instruct models and Thinking models. We categorize the typical mistakes into several types, ranging from perceptual errors to logical inconsistencies in action sequences.

C.1 Case Studies

We list several representative cases below. For each case, we provide the image, the model's response, and a brief analysis of why the model failed to correctly interpret the counter-intuitive action. Specific error cases are presented in Tables 23 to 26. Specifically, Table 23 illustrates the *Statistical Frequency Bias* error. Table 24 demonstrates *Logical/Physical Impossibility*. Table 25 shows *Spatial/Semantic Binding Failure*, and Table 26 presents an example of *Alignment/Extraction Failure*.

Criterion	Pass/Fail Case	Selection / Exclusion Logic
Semantic Consistency	salute/venerate	"Salute" is defined by specific, identifiable hand and eye movements. In contrast, abstract verbs like "venerate" lack a uniform visual manifestation, leading to low annotator agreement on the specific physical action.
Semantic Non-interchangeability	chase/play with	"Play with" is semantically symmetrical; swapping the agent and patient does not fundamentally change the scene's logic. We prioritize asymmetric verbs like "chase," where role reversal creates a distinct counter-intuitive conflict.
Static Recognizability	kick/think	The action must be discernible from a single static frame without temporal context. Cognitive states like "think" are internal and lack observable physical interaction markers, making them unsuitable for VQA tasks.
Commonality	push/block	We select high-frequency, general-purpose verbs like "push" to ensure model evaluation focuses on logic rather than vocabulary. Specialized terms like "block" (e.g., in sports contexts) are excluded due to their heavy reliance on domain-specific knowledge.

Table 10: Detailed Selection and Exclusion Criteria for the Sensitive-Action Verb Library. This taxonomy ensures that all verbs used for counter-intuitive scene generation possess objective visual markers and clear logical asymmetry.

Annotator	Age	Description	Acc	F1	Avg Time (s)
A1	15	Undergraduate student	0.945	0.945	3.7
A2	21	Master student	0.942	0.942	3.3
A3	22	Master student	0.980	0.980	3.4
A4	59	General public	0.953	0.953	4.1
A5	58	General public	0.920	0.920	4.2
Mean	–	–	0.948	0.948	3.740
Std	–	–	0.020	0.020	0.404
Inter-annotator agreement (Fleiss' κ)				0.88	

Table 11: Human annotation statistics on the CAIT VQA task.

Which of the following option better describes the image?
Option A: {option_a}
Option B: {option_b}

Important Note: Answer directly with the option letter (A or B) only.

Table 12: Original Prompt for MLLMs during inference.

Which of the following option better describes the image?
Option A: {option_a}
Option B: {option_b}

Please begin with a brief analysis...

Important Note: You must write...

Table 13: CoT instructions for Instruct Models.

1. Description: Describe the content...
The primary objects or scene in the image.
Visual features such as colors, shapes, and textures.
Any text, symbols, or logos.
The overall layout and composition of the image.

2. Options: Analyze the meaning...
Option A: {option_a}
Option B: {option_b}

3. Reasoning: Compare the image content...
Compare the image content with the description in Option B...
Analyze which option describes the image content more accurately.
Consider the possibility of uncommon...

4. Conclusion: Based on your analysis...
Summarize your reasoning process.
Clearly state the reasons for choosing A or B.
Provide the final answer.

Important Note: At the very end...

Table 14: Detailed Structured Prompt for System 2 thinking mode.

Model	Gemini-2.5-pro	Claude-4.6-Opus	GPT-5.4
GT Prompt (upper bound)	0.983	0.985	0.985
Proprietary Models			
Gemini-2.5-pro	0.838	0.823	0.843
Claude-4.6-Opus	0.803	0.770	0.790
Open-source Instruct MLLMs			
Qwen3-VL-Instruct-8B	**0.773**	**0.758**	**0.755**
Qwen3-VL-Instruct-4B	0.728	0.690	0.703
Qwen2.5-VL-7B	0.668	0.653	0.660
LLaMA3.2-Vision-11B	0.640	0.588	0.593
Gemma-3-12B	0.618	0.585	0.593
Qwen2.5-VL-3B	0.590	0.570	0.593
LLaVA-NeXT-7B	0.558	0.505	0.582
Gemma-3-4B	0.553	0.538	0.532
Open-source Thinking MLLMs			
GLM-4.1V-Thinking-9B	**0.738**	**0.718**	**0.743**
Qwen3-VL-Thinking-8B	0.705	0.678	0.705
Kimi-VL-A3B-Thinking-16B	0.705	0.673	0.690
Qwen3-VL-Thinking-4B	0.693	0.683	0.710
InternVL-3.5-4B	0.688	0.663	0.698
InternVL-3.5-8B	0.683	0.620	0.663

Table 15: Different Judges of LLM-re-Eval. Values denoted in **bold** indicate the highest performance in each open-source model group for each judge.

Model	Param	A-A	B-S	H-A	H-H
∼10B Instruct Models					
Qwen2.5-VL-Instruct	7B	0.486	0.447	0.483	0.352
Qwen3-VL-Instruct	8B	0.500	0.342	0.421	0.297
Gemma-3-it	12B	0.389	0.263	0.434	0.290
LLaMA3.2-Vision	11B	0.389	0.500	0.400	0.310
LLaVA-NeXT	7B	0.556	0.579	0.469	0.414
∼3B Instruct Models					
Qwen2.5-VL-Instruct	3B	0.472	0.289	0.407	0.310
Qwen3-VL-Instruct	4B	0.500	0.263	0.317	0.269
Gemma-3-it	4B	0.500	0.184	0.352	0.290
∼10B Thinking Models					
Qwen3-VL-thinking	8B	0.653	0.579	0.641	0.648
GLM-4.1V-Thinking	9B	0.653	0.579	0.600	0.641
Kimi-VL-A3B-Thinking	16B	0.722	0.500	0.710	0.683
Internvl-3.5	8B	0.403	0.184	0.393	0.283
∼3B Thinking Models					
Qwen3-VL-thinking	4B	0.611	0.684	0.600	0.641
Internvl-3.5	4B	0.639	0.684	0.552	0.655

Table 16: Detailed performance of MLLMs across different domains (A-A: Animal-Animal, B-S: Bio-StillLife, H-A: Human-Animal, H-H: Human-Human).

Model	Param	Acc	F1 score
∼10B Instruct MLLMs			
Qwen2.5-VL-Instruct	7B	0.538	0.534
Qwen3-VL-Instruct	8B	0.750	0.750
Gemma-3-it	12B	0.535	0.538
LLama3.2-vision	11B	0.458	0.458
LLaVA-NeXT	7B	0.490	0.450
∼3B Instruct MLLMs			
Qwen2.5-VL-Instruct	3B	0.413	0.413
Qwen3-VL-Instruct	4B	0.663	0.663
Gemma-3-it	4B	0.473	0.440

Table 17: Performance of Instruct MLLMs using structured prompt.

Model	Param	Acc	F1 score
∼10B Thinking MLLMs			
Qwen3-vl-thinking	8B	0.703	0.703
GLM-4.1V-Thinking	9B	0.675	0.698
Kimi-VL-A3B-Thinking	16B	0.733	0.733
InternVL-3.5	8B	0.610	0.610
∼3B Thinking MLLMs			
Qwen3-vl-thinking	4B	0.720	0.720
InternVL-3.5	4B	0.580	0.575

Table 18: Performance of Thinking MLLMs using structured prompt.

Step 1: Semantic Scene Generation (Verb → Text)

Model: GPT-4o-mini **Temperature:** 0.7

System Prompt:

Please note the following examples I have provided. Each group is divided into two parts: the normal scene and the scene that is logical but extremely uncommon after the master-guest inversion. When reversing, try to keep the general structure of the original sentence's language description unchanged, only swapping the positions of the subject and object. Following these examples, generate semantic scenarios that meet the requirements for the verbs I gave you, with 10 groups for each verb.

(Few-shot examples omitted)

You MUST respond in valid json format. Output a json object with the following structure: `{ "verb": "<input_verb>", "scene_sets": [ { "id": <number>, "common": ["<common_sentence>"], "uncommon": ["<uncommon_sentence>"] } ] }`

User Prompt:

`{ "verb": "<verb>" }`

Output:

```
{ "verb": "<verb>",
  "scene_sets": [ { "id": 1, "common":
["<common_sentence>"], "uncommon":
["<uncommon_sentence>"] }, ... ] }
```

Table 19: Prompt details for Step 1: Generating semantic scene pairs from verbs.

Step 2: Human Verification

Operator: Human Expert **Interface:** Custom GUI

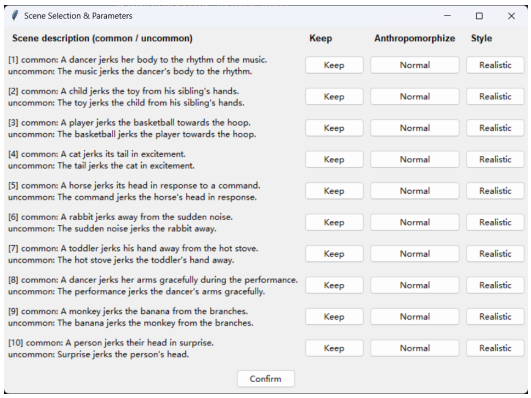

Figure A: The custom GUI for manual scene selection and parameter configuration.

Process Description:

The raw semantic pairs generated in Step 1 are presented to a human annotator via a graphical user interface. The annotator performs quality control and parameter setting for each scene to ensure high usability for image generation.

Human Actions:

1. **Selection:** Review 10 generated scene pairs. Toggle the "Keep" button for scenes that maintain logical causality after inversion and filter out nonsensical ones.
2. **Anthropomorphize:** Set the subject's nature (e.g., "Normal" vs. "Anthropomorphic") based on whether the inverted subject requires human-like traits to perform the action.
3. **Style Setting:** Assign a visual style ("Realistic" or "Cartoon") to guide the subsequent image generation.

Input Data:

`{ "verb": "<verb>", "scene_sets": [ ... (10 raw pairs) ... ] }`

Output Data:

```
{ "verb": "<verb>",
  "scene_sets": [ { "id": 1, "common": [...],
"uncommon": [...], "selected": true,
"anthropomorphize": true/false, "style":
"realistic/cartoon" }, ... ] }
```

Table 20: Details for Step 2: Human-in-the-loop selection and parameter configuration.

Step 3: Scene Description Generation (Text → Description)

Model: GPT-4o-mini **Temperature:** 0.7

System Prompt:

You are given: - verb: {verb}, - scene_sets: {scene_sets_json}

Task: For each uncommon semantic scenario (subject–object inversion) in the scene_sets, generate ONE highly detailed and visually grounded description suitable for Gemini's text-to-image model.
Each description must: 1. Clearly convey the subject–object role reversal described in the uncommon_text. 2. Respect the per-scene options (anthropomorphize, style). 3. Be vivid and concrete, describing what the viewer can see.

Output strictly in json format. You MUST return a valid json object with the following structure: { "verb": "<verb>", "records": [{ "scene_id": 1, "uncommon_image_description": "<description>", ... }] }

User Prompt:

{ "verb": "<verb>", "scene_sets": [{ "id": 1, "uncommon": ["<uncommon_sentence>"], ... }, ...] }

Output:

{ "verb": "<verb>", "records": [{ "scene_id": 1, "uncommon_text": "<uncommon_sentence>", "uncommon_image_description": "<uncommon_image_description>", "style": "realistic/cartoon" }, ...] }

Table 21: Prompt details for Step 3: Converting text pairs into detailed visual descriptions.

Step 4: Paradoxical Image Generation (Description → Image)

Model: Gemini-2.5-flash-image **Temperature:** 0.7

Full Prompt Template:

You are a visual reasoning model that deliberately violates common sense. You must prioritize paradox, contradiction, and reversed causality over realism.

World rules: 1. Physical laws can be inverted. 2. Causes may appear after results. 3. Object scale, weight, and function may contradict logic. 4. The image must look realistic in style but impossible in logic. 5. Do not explain the paradox visually – let it appear naturally.

Here is the uncommon and counter-intuitive scene you have to generate:
<uncommon_image_description>

Image size: 512×512 pixels

Output:

[Generated Image: A 512×512 pixel paradoxical image visualizing the description]

Table 22: Prompt details for Step 4: Generating paradoxical images from descriptions.

CASE 1: Prior Bias Overriding Visual Evidence (Thinking Model - InternVL-3.5-8B)

QUESTION:

Choose the description that best matches the image:
A. a little boy is guiding a parent
B. a parent is guiding a little boy

MODEL THINKING:

Okay, let's tackle this step by step. First, I need to analyze the image as per the instructions. The image shows a woman and a young boy holding hands and running down a tree-lined path. The woman is smiling broadly, and the boy looks happy too. [...]

The key here is to determine who is leading whom. The woman is holding the boy's hand, and they're both moving forward. The boy is slightly ahead, but the woman is supporting him by holding his hand. **Typically, when a parent guides a child, they might hold the child's hand and lead them, especially if the child is running.** The boy's position isn't necessarily leading the woman; instead, they seem to be moving together with the woman supporting him. So, the parent (woman) is guiding the child (boy) as they both move forward. That would make option B more accurate. However, sometimes the child might take the lead, but in this case, the woman's posture and the way they're holding hands suggest she's guiding him. **The boy is smiling and seems to be enjoying the run, but the guiding action is more about the parent leading the child.** So, I think B is the correct answer here.

Final Answer: B

FAILURE ANALYSIS:

Acknowledgement of Evidence: The model initially acknowledges the valid visual cue, noting that "The boy is slightly ahead."

Activation of Priors: However, it immediately counters this observation with a probabilistic generalization derived from its training data: "Typically, when a parent guides a child... they lead them."

Rejection of Reality: Driven by this statistical likelihood, the model rejects the low-probability visual reality (a child guiding a parent). It rationalizes the scene to fit the high-probability distribution, arguing the boy is "not necessarily leading," effectively hallucinating a scenario that aligns with its internal priors rather than the image.

Table 23: Error Analysis Case 1: Type 1 – Statistical Frequency Bias

CASE 2: Logical Impossibility (Instruction Model - Gemma3-12B-it)

QUESTION:

Choose the description that best matches the image:
A. A cat swats a toy mouse.
B. A toy mouse swats the cat.

REASONING TRACE:

Okay, let's analyze the image and the provided options... The image depicts a domestic cat and a toy mouse...

Let's compare the image content with each option... Option A: The image clearly shows the cat with its paw extended... The cat is the active agent...

Option B: The image does *not* depict the toy mouse striking the cat. **The toy mouse is stationary and passive. The concept of a toy mouse "swatting" a cat is physically impossible and defies the typical understanding of the relationship between a cat and its toys.** It's an unconventional and highly implausible scenario...

My reasoning process involved a detailed description of the image, followed by a comparison of the image content with each option. Option A, "A cat swats a toy mouse," accurately reflects the scene, as the cat is actively engaging with the toy. **Option B is simply not possible... Therefore, Option A is the correct choice.**

Final Answer: A

FAILURE ANALYSIS:

Acknowledgement of Entities: The model correctly identifies the scene's components ("domestic cat," "toy mouse") and their spatial proximity.

Activation of Constraints: However, it invokes a strict logical rule regarding inanimate objects, explicitly labeling the scenario as "physically impossible" and "highly implausible" because toys cannot initiate action.

Rejection of Reality: Instead of accepting the staged or humorous nature of the image (where a toy might visually "swat" a cat), the model rigidly rejects the visual evidence. It forces the interpretation to align with real-world physics, concluding that the event is "simply not possible" despite the visual cue.

Table 24: Error Analysis Case 2: Type 2 – Logical Impossibility

CASE 3: Semantic Binding Failure (Instruct Model - Gemma3-12B-it)

QUESTION:

Choose the description that best matches the image:
A. A knight stabs a dragon with a sword.
B. A dragon stabs the knight with a sword.

REASONING TRACE:

Okay, let's analyze the image... The image depicts a dramatic, fantastical scene of a knight battling a large, red dragon...

Let's compare the image content with each option... Option A (Knight stabs dragon): The knight *is* holding a sword raised, and it appears to be aimed at the dragon. The sword is close to the dragon's body. The angle of the sword and the knight's posture suggest an attack. This aligns well with the image.

Option B (Dragon stabs knight): **The dragon is holding a sword, but the sword is pointed downwards, away from the knight... there's no indication that the dragon is actively stabbing him.** The dragon's posture is more of a defensive or aggressive stance, rather than a stabbing motion. This does not align well with the image.

While it's possible to interpret the scene in various ways, the most straightforward interpretation is that the knight is attempting to strike the dragon. **Therefore, the image more accurately represents a knight attempting to stab a dragon.**

Final Answer: A

FAILURE ANALYSIS:

Object Identification: The model correctly detects all key elements in the scene: the knight, the dragon, and the swords.

Attribute Misbinding: However, it fails to correctly bind the action to the agent. It misinterprets the spatial orientation of the dragon's sword (acknowledging it is "pointed downwards" but failing to recognize the attack) while incorrectly binding the offensive action to the knight's posture, effectively ignoring the sword in the dragon's hand.

Subject-Object Reversal: Consequently, the model swaps the semantic roles of the entities. It assigns the "agent" role (stabbing) to the knight and the "patient" role to the dragon, directly contradicting the visual evidence where the dragon is the aggressor.

Table 25: Error Analysis Case 3: Type 3 – Semantic Binding Failure

CASE 4: Alignment Failure (Thinking Model - InternVL-3.5-8B)

QUESTION:

Choose the description that best matches the image:
A. a father is guiding a girl riding a bike on the road
B. a girl on the road is guiding her father riding a bike

MODEL THINKING:

Okay, let's tackle this step by step. [...] The scene shows a man riding a bicycle on a residential street, and a young girl walking beside him. The girl is holding a rope attached to the bike's handlebars. [...] Option A says a father is guiding a girl riding a bike. **But in the image, the girl isn't on the bike; she's walking alongside.** So Option A might be incorrect. [...] The key is who is riding the bike. In the image, the father is on the bike. Therefore, Option A is incorrect because the girl isn't on the bike. Option B is the only remaining choice, even if it's unconventional. [...] **Therefore, the correct answer must be Option B, where the girl is guiding her father who is riding the bike.** Even though it's unconventional, the image shows the girl holding the rope attached to the bike's handlebars, which could imply she's steering it, thus guiding the father.

Final Answer: A

FAILURE ANALYSIS:

Correct Visual Reasoning: Unlike other types of errors, the model correctly perceives the visual reality in its chain-of-thought, explicitly stating "the girl isn't on the bike" and identifying that Option A is factually incorrect.

Final Answer Generation Collapse: Despite deducing the correct answer (B) internally, the model fails at the final generation step. The explicit reasoning is discarded in favor of a hallucinated label.

Re-emergence of Priors: This disconnect highlights a specific alignment failure. The strong textual prior associated with "father guiding girl" (common) versus "girl guiding father" (rare) overwhelms the logical conclusion during the token decoding phase, causing the model to revert to the statistical mean despite its own correct reasoning.

Table 26: Error Analysis Case 4: Type 4 – Alignment Failure