MuSLR: Multimodal Symbolic Logical Reasoning

Jundong Xu[1], **Hao Fei**[1]*, **Yuhui Zhang**[2], **Liangming Pan**[3], **Qijun Huang**[4], **Qian Liu**[5],
Preslav Nakov[6], **Min-Yen Kan**[1], **William Yang Wang**[7], **Mong-Li Lee**[1], **Wynne Hsu**[1]
[1] National University of Singapore, [2] Stanford University, [3] Peking University, [4] UniMelb,
[5] University of Auckland, [6] MBZUAI, [7] University of California, Santa Barbara
`jundong.xu@u.nus.edu`, `haofei37@nus.edu.sg`, `yuhuiz@cs.stanford.edu`,
`liangmingpan@pku.edu.cn`, `qijunhuang@student.unimelb.edu.au`,
`liu.qian@auckland.ac.nz`, `preslav.nakov@mbzuai.ac.ae`, `knmnyn@nus.edu.sg`,
`william@cs.ucsb.edu`, `dcsleeml@nus.edu.sg`, `dcshsuw@nus.edu.sg`

Abstract

Multimodal symbolic logical reasoning, which aims to deduce new facts from multimodal input via formal logic, is critical in high-stakes applications such as autonomous driving and medical diagnosis, as its rigorous, deterministic reasoning helps prevent serious consequences. To evaluate such capabilities of current state-of-the-art vision language models (VLMs), we introduce **MuSLR**, the first multimodal symbolic logical reasoning grounded in formal logical rules. We curate a benchmark dataset for MuSLR comprising 1,093 instances across 7 domains, including 35 atomic symbolic logic and 976 logical combinations, with reasoning depths ranging from 2 to 9. We evaluate 7 state-of-the-art VLMs on our benchmark and find that they all struggle with multimodal symbolic reasoning, with the best model, GPT-4.1, achieving only 46.8%. Thus, we propose **LogiCAM**, a modular framework that applies formal logical rules to multimodal inputs, boosting GPT-4.1's Chain-of-Thought performance by 14.13%, and delivering even larger gains on complex logics such as first-order logic. We also conduct a comprehensive error analysis, showing that around 70% of failures stem from logical misalignment between modalities, offering key insights to guide future improvements. All data and code are publicly available at `https://llm-symbol.github.io/MuSLR`.

1 Introduction

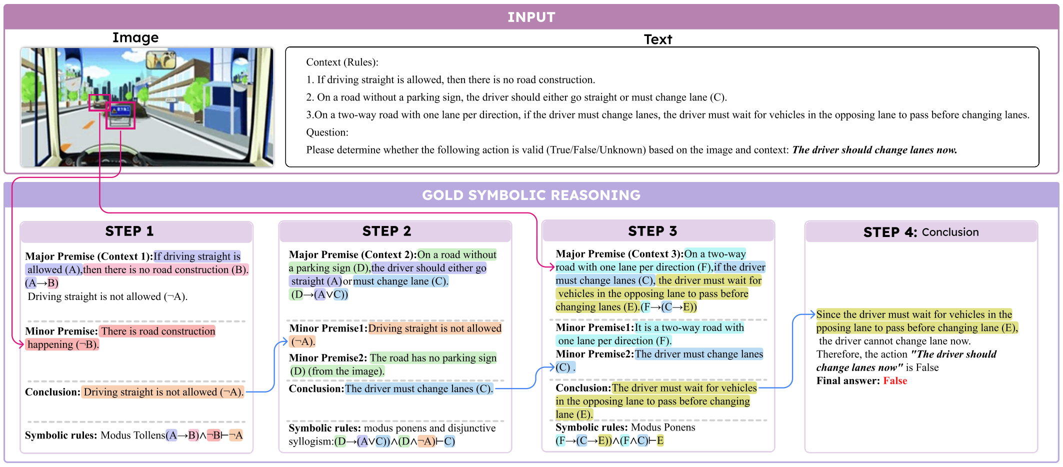
Recent progress has extensively highlighted the pivotal role of reasoning capabilities in enhancing the generality and robustness of large language models (LLMs) [9, 12, 29–31]. Yet, achieving human-level intelligence demands more than commonsense or heuristic thinking. In particular, *symbolic logical reasoning*, grounded in formal logic such as first-order logic, offers a rigorous, precise, and verifiable paradigm essential for high-stakes scenarios where reasoning errors can have critical consequences. Although previous works have shown that LLMs can handle symbolic reasoning in purely textual contexts [22, 36, 37], these capabilities remain limited to unimodal inputs, i.e., text. However, many real-world domains, such as autonomous driving, healthcare, law, and finance, demand reasoning that integrates multiple modalities, particularly combining visual and textual information, to support accurate and reliable conclusions. Consider an autonomous driving system that observes a traffic sign (from a camera image) indicating "Road Closed Ahead", given the traffic rule "*Only if the road ahead is open (B), the vehicles may proceed straight (A).*" From the image, the system detects that the road is in fact closed ($\neg B$), and must infer that continuing straight is not permitted ($\neg A$), forming a formal logical reasoning (Modus Tollens; $(A \to B) \land \neg B \to \neg A$) to avoid traffic accidents. Despite the significance of such multimodal symbolic reasoning, no standard definition or benchmark currently exists for this capability.

To fill this gap, we introduce ***Mu**ltimodal **S**ymbolic **L**ogical **R**easoning* (**MuSLR**), a novel task that challenges VLMs to perform symbolic reasoning over combined visual and textual inputs. Figure 1 illustrates the MuSLR task with the above example. We define MuSLR under two task formats: *Truth*

---

[*] Corresponding author: Hao Fei

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



Context (Rules):

1. If driving straight is allowed, then there is no road construction.
2. On a road without a parking sign, the driver should either go straight or must change lane (C).
3. On a two-way road with one lane per direction, if the driver must change lanes, the driver must wait for vehicles in the opposing lane to pass before changing lanes.

Question:

Please determine whether the following action is valid (True/False/Unknown) based on the image and context: ***The driver should change lanes now.***

Major Premise (Context 1): If driving straight is allowed (A), then there is no road construction (B). $(A \rightarrow B)$ Driving straight is not allowed ($\neg A$).

Minor Premise: There is road construction happening ($\neg B$).

Conclusion: Driving straight is not allowed ($\neg A$).

Symbolic rules: Modus Tollens $(A \rightarrow B) \wedge \neg B \vdash \neg A$

Major Premise (Context 2): On a road without a parking sign (D), the driver should either go straight (A) or must change lane (C). $(D \rightarrow (A \vee C))$

Minor Premise1: Driving straight is not allowed ($\neg A$).

Minor Premise2: The road has no parking sign (D) (from the image).

Conclusion: The driver must change lanes (C).

Symbolic rules: modus ponens and disjunctive syllogism: $(D \rightarrow (A \vee C)) \wedge (D \wedge \neg A) \vdash C)$

Major Premise (Context 3): On a two-way road with one lane per direction (F), if the driver must change lanes (C), the driver must wait for vehicles in the opposing lane to pass before changing lanes (E). $(F \rightarrow (C \rightarrow E))$

Minor Premise1: It is a two-way road with one lane per direction (F).

Minor Premise2: The driver must change lanes (C) .

Conclusion: The driver must wait for vehicles in the opposing lane to pass before changing lane (E).

Symbolic rules: Modus Ponens $(F \rightarrow (C \rightarrow E)) \wedge (F \wedge C) \vdash E$

Since the driver must wait for vehicles in the pposing lane to pass before changing lane (E), the driver cannot change lane now. Therefore, the action ***"The driver should change lanes now"*** is False

Final answer: **False**

Figure 1: An example of a depth-4 propositional logic task, requiring the VLMs to apply formal symbolic logic rules and integrate multimodalities to reach the conclusion.

Evaluation and *Multiple Choice*, where given an image I, context T, the model must apply symbolic logical reasoning to identify the correct answer. To enable systematic evaluation, we then propose **MuSLR-Bench**, a high-quality benchmark dataset specifically designed to assess the symbolic reasoning abilities of state-of-the-art VLMs. Drawing from authentic web-sourced scenarios where visual and textual content naturally co-occur, we annotate each instance with formal logical rules (e.g., modus ponens) and conduct rigorous quality checks to ensure correctness and logical validity. MuSLR-Bench comprises 1,093 instances spanning 7 domains, including 35 atomic symbolic logic and 976 complex logical compositions, with reasoning depths ranging from 2 to 9 to reflect diverse difficulty levels. In a pilot study, we evaluate seven leading open- and closed-weight VLMs of varying sizes on MuSLR-Bench, revealing that even top models struggle substantially with multimodal symbolic logic inference.

To establish a strong baseline for MuSLR, we further propose **LogiCAM** (**Logi**cal reasoning with **C**ommonsense **A**ugmentation in **M**ultimodalities), which decomposes multimodal symbolic reasoning into modular steps through Chain-of-Thought (CoT) mechanism (cf. Figure 4). First, the `Premise Selector` is designed to address the difficulty of multimodal fusion. We next devise a `Reasoner` module to integrate multimodal evidence and apply symbolic reasoning by approximating formal logical rules, enabling rigorous and systematic deduction to meet the core challenge of MuSLR. Then, the `Reasoning Type Identifier` is designed to address the issue of incomplete information in MuSLR, where heuristics act as supplementary resources to complement symbolic rules when they are insufficient to reach the conclusion. Extensive experiments show that LogiCAM improves GPT-4.1's CoT performance by 14.13% on MuSLR-Bench, achieving even greater gains on complex first-order logic tasks. Further analysis reveals that reasoning performance deteriorates sharply as logical complexity and chain depth increase, highlighting key limitations of current popular VLMs.

In summary, our contributions are fourfold:

- We introduce **MuSLR**, a pioneering task targeting multimodal symbolic logical reasoning, addressing a critical gap in real-world AI reasoning.
- We curate **MuSLR-Bench**, a high-quality dataset comprising 1,093 instances with diverse logical structures and depths, serving as a critical foundation for this topic.
- We develop **LogiCAM**, a strong CoT-based baseline method that decomposes complex reasoning into more manageable and trackable modules.
- Through extensive experiments and analyses, we pinpoint where and why current VLMs struggle with MuSLR, offering insights for future investigation of this area.

2 Related Work

Textual Symbolic Logic Reasoning and Benchmarks. Existing benchmarks for symbolic logical reasoning have primarily focused on purely textual settings under formal logic rules. For instance, FOLIO [7] is a human-annotated dataset for complex natural language reasoning equipped with first-order logic annotations to ensure the logical consistency of premises and conclusions. ProofWriter [28] provides small English rulebases of facts and rules with associated questions, requiring models

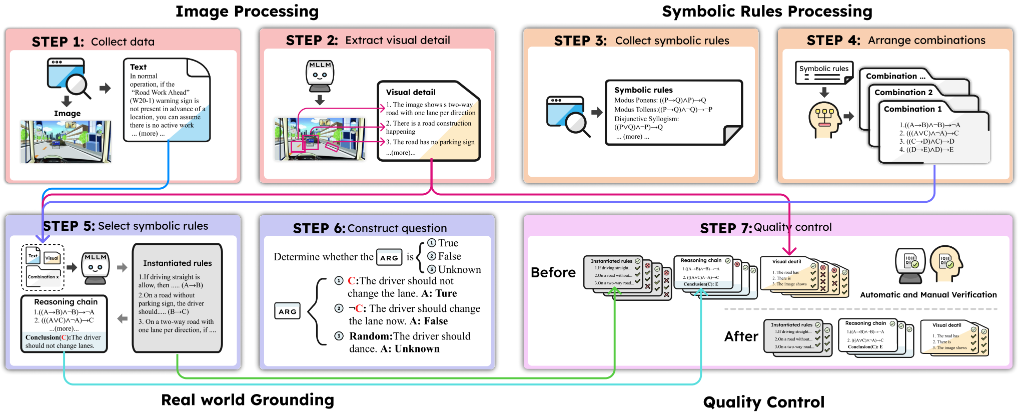


Figure 2: Pipeline of MuSLR data construction. We begin by collecting multimodal data and symbolic rules. These rules are then combined to form reasoning chains, which are grounded in real-world contexts to generate questions and answers, followed by a strict quality check.

to prove or refute statements (or answer "unknown" when proof is impossible) via multi-step natural language proofs. Likewise, Multi-LogiEval [24] evaluates multi-step logical reasoning across propositional, first-order, and even non-monotonic logic types, encompassing over 30 inference rules and various depths to test LLMs' deductive abilities. We further acknowledge numerous additional related works, such as ProntoQA [27], LogicBench [23], and RuleArena [41]. However, these benchmarks assume fully specified, idealized inputs in a single modality (text) and do not incorporate visual information, limiting their direct applicability to real-world scenarios.

Multimodal Reasoning and Benchmarks. In parallel, several benchmarks have introduced accessing reasoning in vision and language [5, 33]. LogicVista [35] evaluates VLMs' logical reasoning in visual contexts, with 448 annotated multiple-choice questions spanning a spectrum of logical reasoning tasks and capabilities. Similarly, VisuLogic [38] targets vision-centric reasoning by constructing tasks that require robust visual logic without relying on textual descriptions or shortcuts. Meanwhile, broader vision-language benchmarks emphasize contextual reasoning rather than formal logic: for example, MMMU [39] offers college-level multimodal questions across six disciplines (e.g., charts, maps, chemical structures), testing domain-expert reasoning. MathVista [18] targets compositional mathematical inference in visual scenarios. However, none of these multimodal benchmarks explicitly test the application of formal logical rules (e.g. Modus Ponens or De Morgan's Law) grounded in both visual and textual input. **MuSLR** addresses this gap by requiring explicit symbolic logical deduction from joint visual–textual inputs, integrating formal logic rules into multimodal understanding.

Neuro-Symbolic Reasoning Method. Many prior works adopt a symbolic prover in the reasoning pipeline to achieve rigorous and reliable reasoning. Typically, an LLM is used to formalize natural language into symbolic form, after which a theorem prover is employed to solve it [11, 19, 22, 26, 34]. However, theorem provers only accept text input. In multimodal scenarios, this requires first converting visual or multimodal information into text, a process that inevitably leads to information loss and thus limits adaptability. In contrast, our LogiCAM framework is designed to approximate symbolic reasoning using a vision–language model (VLM), which has direct access to multimodal information without relying on lossy translation.

3 Task Definition

The proposed tasks require models to integrate information from both an image I and a text passage T to perform reasoning, ensuring that neither modality alone is sufficient for correct inference. The tasks explicitly emphasize **multimodal reasoning**, where the fusion of visual and textual context is essential for deriving accurate and consistent conclusions.

Task-I: Truth Evaluation (True/False/Unknown) Question. Given an image I, a text passage T, and an argument A, the model must determine the truth value of the argument based on the combined information from I and T. Specifically, the model outputs the truth value $\text{Truth}(A) \in \{\text{True}, \text{False}, \text{Unknown}\}$ and generates a sequence of reasoning steps $R = \{R_1, R_2, \dots, R_n\}$, where each R_i represents an individual step that contributes to the final decision. Formally, the input is a triplet (I, T, A), and the output consists of $\text{Truth}(A)$ and R.

Statistics	Numbers
Total instances	1093
Total sources	7
Domain (#)	7
Symbolic logic (#)	3
Atomic symbolic rules (#)	35
Symbolic rule combination (#)	976
Min reasoning depth	2
Max reasoning depth	9
Min context length	35
Max context length	1484
Avg. context length	554.9

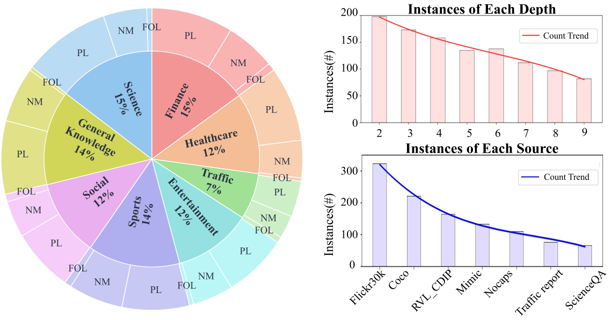


Figure 3: Dataset Statistics. The left table presents general dataset statistics. The middle pie chart illustrates the distribution across domains and symbolic logic. The right bar charts display the number of instances by reasoning depth and data source.

Task-II: Multiple Choice Question. Given an image I, a text passage T, and candidate arguments $\{A_1, A_2, A_3, A_4\}$, the model must select the argument that best matches the image and text, denoted as $\text{BestArgument}(I, T) \in \{A_1, A_2, A_3, A_4\}$. Additionally, the model must provide detailed reasoning steps $R = \{R_1, R_2, \ldots, R_n\}$, where each R_i details a step in the reasoning process. Formally, the input is a triplet $(I, T, \{A_1, A_2, A_3, A_4\})$, and the output consists of $\text{BestArgument}(I, T)$ and R.

4 MuSLR-Bench: A Benchmark for Multimodal Symbolic Logical Reasoning

Dataset Construction. We collect images from various sources such as COCO [14], Flickr30k [25], nocaps [1], Mimic [10], RVL_CDIP [8], ScienceQA [17], and manually collected Traffic Reports. Visual details for each image are extracted using GPT-4o, ensuring diverse and fine-grained descriptions. We carefully select non-trivial logical inference rules, such as Modus Ponens and Hypothetical Syllogism, drawn from propositional logic (PL), first-order logic (FOL), and non-monotonic logic (NM). These rules then form meaningful but abstract reasoning chains through logical combinations. The abstract chains are grounded in real-world contexts by leveraging extracted visual features and relevant retrieved text from sources like healthcare, traffic reports, and Wikipedia. Questions and answers are then generated based on these instantiated reasoning chains, using rule-based substitution.

To ensure the quality and relevance of the dataset, both automatic and manual quality control procedures are employed. Automatic checks include assessing lexical similarity and commonsense plausibility, while human annotators verify the accuracy of visual details and the real-world relevance of the generated context. Instances that fail these checks are filtered out, ensuring a high-quality, logically sound, and contextually relevant dataset. Further details on the data construction and quality control processes are provided in the Appendix B and C, respectively.

4.1 Dataset Highlights

MuSLR consists of 1093 instances, where each instance includes a multimodal context (image and associated text), a ground-truth logical reasoning chain, and corresponding question-answer pairs. The dataset is constructed to support both detailed symbolic logical reasoning analysis and challenging multimodal reasoning tasks. Below, we summarize the key features of the dataset:

Ground-Truth Reasoning Steps. Each instance is equipped with an explicit, step-by-step ground-truth reasoning chain, enabling detailed analysis and training of models for symbolic logical reasoning.

Multi-Scenario Coverage. The dataset spans a wide range of domains, including science, entertainment, sports, social issues, general knowledge, traffic, healthcare, and finance. The distribution across these scenarios is illustrated in the pie chart in Figure 3.

Diverse Symbolic Reasoning Types. MuSLR contains diverse symbolic logic: propositional logic (PL), first-order logic (FOL), and non-monotonic logic (NM), ensuring broad logical coverage.

Multimodality. To the best of our knowledge, this is the first dataset that combines both image and text modalities for symbolic logical reasoning tasks grounded in formal logical rules.

Diverse Difficulty Levels. The reasoning chains vary in depth from 2 to 9 steps, offering a broad spectrum of difficulty levels and supporting evaluation across simple and complex reasoning scenarios.

Multiple Question Types. The dataset supports multiple question formats, including **Truth Evaluation** and **Multiple-Choice** questions, allowing for diverse model evaluation protocols.

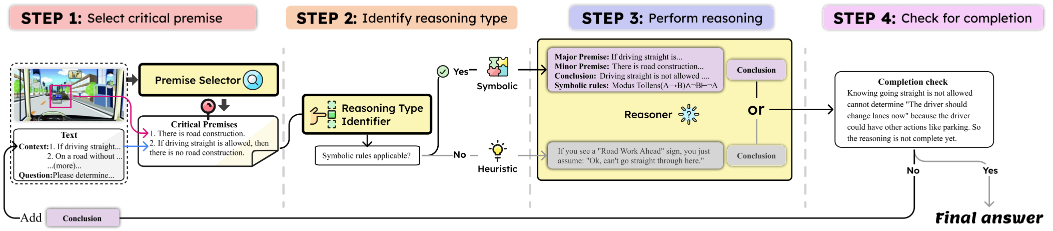

STEP 1: Select critical premise
STEP 2: Identify reasoning type
STEP 3: Perform reasoning
STEP 4: Check for completion
Premise Selector
Text
Context:1. If driving straight...
2. On a road without ...
...(more)...
Question:Please determine...
Critical Premises
1. There is road construction.
2. If driving straight is allowed, then there is no road construction.
Reasoning Type Identifier
Symbolic rules applicable?
Yes
No
Symbolic
Heuristic
Major Premise: If driving straight is...
Minor Premise: There is road construction...
Conclusion: Driving straight is not allowed ....
Symbolic rules: Modus Tollens(A→B)∧¬B⊢¬A
Conclusion
Reasoner
or
If you see a "Road Work Ahead" sign, you just assume: "Ok, can't go straight through here."
Conclusion
Completion check
Knowing going straight is not allowed cannot determine "The driver should change lanes now" because the driver could have other actions like parking. So the reasoning is not complete yet.
No
Yes
Final answer
Add
Conclusion

Figure 4: **LogiCAM** Workflow. The figure illustrates a single iteration; the complete multi-iteration reasoning process is detailed in Section 9.

4.2 Challenge

MuSLR presents five key challenges for developing robust multimodal symbolic reasoning models:

Integrate Multimodality. Can the model extract and integrate critical visual and textual context to construct valid reasoning chains? (See Section 7.5)

Step-by-Step Symbolic Reasoning Tracability. Can the model produce interpretable, verifiable, step-by-step reasoning processes in valid logic? (See Section 7.2)

Blend Heuristics for Symbolic Reasoning. Can the model apply heuristic reasoning when symbolic logic is insufficient?

Diverse Symbolic Logic. Can the model handle various forms of symbolic logic (PL, FOL, and NM)? (See Section 7.1)

Reasoning Depth Handling. Can the model reason over different depths, maintaining consistency in longer chains? (See Section 7.3)

Addressing these challenges requires models to integrate multimodal perception and systematic logical reasoning, thereby providing a solid foundation for advancing multimodal reasoning systems.

5 LogiCAM: A Modular MuSLR Framework

We propose a modular framework, **LogiCAM** (**Logi**cal reasoning with **C**ommonsense **A**ugmentation with **M**ultimodality), which consists of three modules based on `GPT-4.1`, as illustrated in Figure 4. Each module is designed to address a specific challenge posed by MuSLR. The modules work together to solve different problem components, which include: (1) the *Premise Selector*, (2) the *Reasoning Type Identifier*, and (3) the *Reasoner* module. Below, we explain how each module addresses its challenge and contributes to the reasoning chain.

Select Critical Multimodal Premises. The *Premise Selector* is designed to address the multimodalities integration challenge, which involves the need to process both visual and textual data to extract critical premises. Given an image I and textual information T containing context $\mathcal{T}$ and question Q, this module directs the VLM to first select the most relevant symbolic rules $R_\text{r} \in \mathcal{T}$. The VLM will then analyze the symbolic logic R_r to determine which part is relevant to the image and extract the corresponding visual information V_r. In this way, the system ensures that only the most critical visual and textual details are extracted, avoiding unnecessary complexity and noise from abundant data. The symbolic rule R_r and visual details V_r will be combined and denoted as I_critical.

Identify Reasoning Type. The *Reasoning Type Identifier* addresses the blend of heuristics and symbolic, which involves determining whether symbolic reasoning or heuristics should be applied during each reasoning iteration. The core challenge is deciding when symbolic logic is sufficient and when heuristics should be used to complement symbolic reasoning. To solve this, the *Reasoning Type Identifier* analyzes the selected premises I_critical and determines whether formal logical rules can be applied. If so, prioritize it. Otherwise, heuristics and commonsense reasoning are employed to compensate for the limitations of purely symbolic reasoning. In this way, the model maximizes the rigor and soundness of the reasoning by prioritizing symbolic reasoning while maintaining flexibility to supplement additional knowledge through commonsense-driven heuristics when symbolic reasoning alone is insufficient.

Perform Reasoning. The *Reasoner* is central to addressing symbolic reasoning tracability, which uses a VLM to approximate formal logical rules when symbolic reasoning is required. Depending on the outcome of the Reasoning Type Identifier, the *Reasoner* either applies symbolic reasoning or uses heuristic commonsense to complete the reasoning process. If *symbolic reasoning* is selected,

Model	**Symbol**	**Healthcare**	**Traffic**	**Sports**	**Ent.**	**Social**	**Science**	**Finance**	**General**
Three-shots CoT Open-Weight VLMs									
Qwen	PL	50.00	33.33	33.33	42.67	36.49	48.54	54.17	46.51
	FOL	0.00	42.86	50.00	40.00	66.67	16.67	22.22	25.00
	NM	43.18	25.93	43.75	35.42	23.26	43.75	54.24	37.50
Llava	PL	20.45	30.95	37.18	32.00	22.97	34.95	27.08	43.02
	FOL	0.00	57.14	50.00	20.00	44.44	66.67	55.56	50.00
	NM	31.82	37.04	45.31	43.75	39.53	47.06	25.42	45.31
InternVL	PL	57.95	42.86	37.97	44.00	37.84	46.60	51.04	50.00
	FOL	50.00	42.86	50.00	20.00	66.67	50.00	22.22	50.00
	NM	38.64	29.63	46.88	35.42	46.51	49.02	45.76	43.08
InstructBlip	PL	42.05	33,33	39.2	26.67	36.49	29.13	40.62	25.58
	FOL	50.00	28.57	25.00	40.00	55.56	16.67	22.22	25.00
	NM	52.27	40.74	53.12	31.25	44.19	35.29	2.34	30.77
Three-shots CoT Closed-Weight VLMs									
Claude	PL	44.32	26.19	24.36	26.67	28.38	35.92	36.46	34.88
	FOL	50.00	14.29	50.00	20.00	55.56	0.00	44.44	75.00
	NM	29.55	37.04	32.81	29.17	30.23	43.14	38.98	31.25
GPT-4o	PL	45.45	40.48	33.33	37.50	34.72	37.00	28.99	43.90
	FOL	0.00	14.29	25.00	0.00	37.50	33.33	50.00	33.33
	NM	52.27	48.15	35.48	50.00	52.38	45.10	41.46	32.81
GPT-4.1	PL	54.55	50.00	44.30	41.33	33.78	43.69	45.83	51.16
	FOL	0.00	14.29	50.00	20.00	44.44	16.67	33.33	0.00
	NM	47.73	59.26	46.88	50.00	53.49	56.86	61.02	40.62
LogiCAM	PL	63.64 (+9.09)	61.90 (+11.90)	58.23 (+13.93)	64.00 (+22.67)	56.76 (+22.98)	57.28 (+13.59)	53.68 (+7.85)	67.44 (+16.28)
	FOL	50.00 (+50.00)	60.42 (+46.13)	50.00 (+0.00)	60.00 (+40.00)	44.44 (+0.00)	40.00 (+23.33)	75.00 (+41.67)	75.00 (+75.00)
	NM	63.64 (+15.91)	66.67 (+7.41)	58.23 (+11.35)	60.42 (+10.42)	74.42 (+20.93)	64.71 (+7.85)	74.14 (+13.12)	55.38 (+14.76)

Table 1: Main Results. Blue indicates the best open-weight VLM, and Red indicates the best closed-weight VLM. The (red brackets) indicate our improvement over the base model.

the module applies formal logical rules to the premises I_{critical} and derives a conclusion C based on a syllogism, which draws a result from the major and minor premises. This reasoning process ensures that conclusions are drawn according to sound logical principles. If *heuristics* are selected, the module uses commonsense reasoning to bridge gaps left by symbolic logic. This design makes sure that the model can perform symbolic reasoning grounded in logical principle, while relax this restriction when heuristics are required. A full example can be found in the Figure 9.

Check for Completion. Finally, the system checks whether the conclusion C is sufficient to answer the question Q. If so, it concludes the final answer. Otherwise, the system appends the conclusion C to the context $\mathcal{T}$, resulting in $T' = T \cup C$, and starts over the whole reasoning iteration.

6 Experiments

6.1 Settings

Evaluation. We evaluate models based on two dimensions: direct answer match and reasoning accuracy. Direct answer match measures the correctness of the final answer, while reasoning accuracy evaluates the quality of the step-by-step reasoning. Reasoning accuracy is computed by comparing model-generated steps with ground-truth steps using ROUGE-L [13] and BertScore-F1 [40]. We also assess ROSCOE [6], which measures logical coherence, factual grounding, and informativeness step by step. More details are in Section 7.2.

Baseline. For benchmarking, we consider multiple state-of-the-art models. For open-source models, we benchmark `Qwen2.5-VL-7B-Instruct` [3], `Llava-1.5-7B` [15], `InternVL3-8B` [42], and `Instructblip-Vicuna-13B` [4]. For closed-source models, we evaluate `GPT-4o` [20], `GPT-4.1` [21] and `Claude-3.7-Sonnet` [2]. These models are chosen to represent the current SoTA in multimodal reasoning.

Settings. To ensure reproducibility, all models are evaluated under standardized settings. We adopt a three-shot Chain-of-Thought (CoT) [32] prompting setup. For language model sampling, the temperature is set to 0.0 to minimize randomness and encourage deterministic outputs.

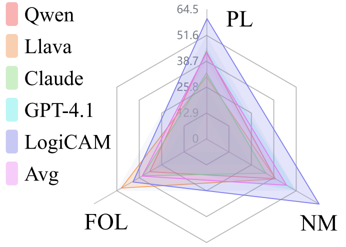


Figure 5: Accuracy of symbolic logic

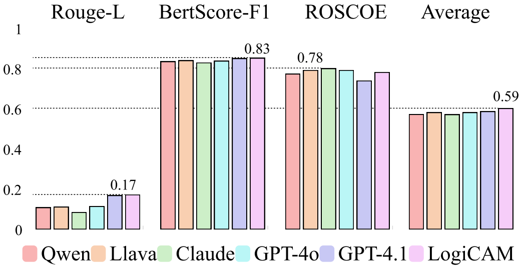


Figure 6: Comparison of models' reasoning tracability

6.2 Main Results and Observations

The main results are presented in Table 1. We have the following observations:

Closed-weight models generally outperform, but open-weight models can rival or surpass them. GPT-4.1 leads with 46.84%, followed closely by InternVL at 45.20%, the top open-weight model. Qwen (41.63%) and GPT-4o (38.93%) follow in the second tier, with InstructBLIP (35.59%), Llava (35.13%), and Claude (33.49%) at the lower end. The performance gap between top and bottom is just 13.35%. These results show that while closed-weight models typically excel, well-designed open-weight models can sometimes outperform proprietary models

LogiCAM enhances CoT and achieves the highest overall performance, with especially strong gains in complex symbolic logic. Integrating LogiCAM into GPT-4.1 results in a substantial performance boost, increasing the average accuracy by 14.13%. When examined by logic type, the improvements are consistent yet differ in scale: FOL accuracy increases by 48.93 %, PL by 31.93 %, and NM by 26.17 %. This pattern indicates that the advantage of LogiCAM grows with the complexity of the logic type: the largest relative improvement is observed in FOL, the most structurally demanding form, followed by PL, and then NM, which is more aligned with intuitive human reasoning and less dependent on rigid symbolic structure. These results suggest that LogiCAM not only strengthens general symbolic reasoning but is especially effective in complex logical operations.

7 Analysis and Discussion

We conduct additional experiments and perform detailed analysis to gain deeper insights into the multimodal symbolic reasoning capabilities of current VLMs.

7.1 Effects on Different Types of Symbolic Logic

In Figure 5, we evaluate the accuracy of each symbolic logic and found that **Model accuracy decreases with rising symbolic complexity: VLMs perform best with non-monotonic reasoning, less well with propositional logic, and struggle most with first-order logic.** First-order logic has the lowest average accuracy at 37.04%, due to its strictest formalism and need for precise variable binding and quantifier tracking. Propositional logic fares better with 42.77%, as its simpler structure eases syntactic constraints. Non-monotonic reasoning performs best at 46.09%, due to its closer alignment with human cognition and requiring less rigid symbolic manipulation. Overall, as symbolic complexity increases, model accuracy declines, highlighting the challenges of fine-grained logical abstraction in current VLMs.

7.2 Tracability of Reasoning Step

As shown in Figure 6, LogiCAM leads in both ROUGE-L (0.170) and BertScore (0.835), with the highest overall mean (0.590), indicating its outputs closely match human phrasing and meaning. Claude scores highest on ROSCOE (0.784), reflecting strong logical consistency but performs poorly on ROUGE-L (0.084). GPT-4.1 balances phrasing and semantics (ROUGE-L = 0.166%, BertScore = 0.833%) but shows moderate stepwise justification (ROSCOE = 0.725%), suggesting occasional logical gaps. Llava and GPT-4o have similar profiles (Average = 0.570%), demonstrating that strong semantic similarity (0.822%) doesn't guarantee superior inference quality (ROSCOE = 0.776%).

Surface-level or semantic objectives alone don't ensure logical coherence. Future work should include logic-focused training goals. A Pearson's correlation analysis reveals a weak correlation between ROUGE-L and ROSCOE (r = 0.25) but a moderate correlation between BertScore and ROSCOE (r = 0.65), suggesting that surface-level metrics do little for logical coherence, while semantically rich training helps more. Claude's high ROSCOE but low ROUGE and BertScore

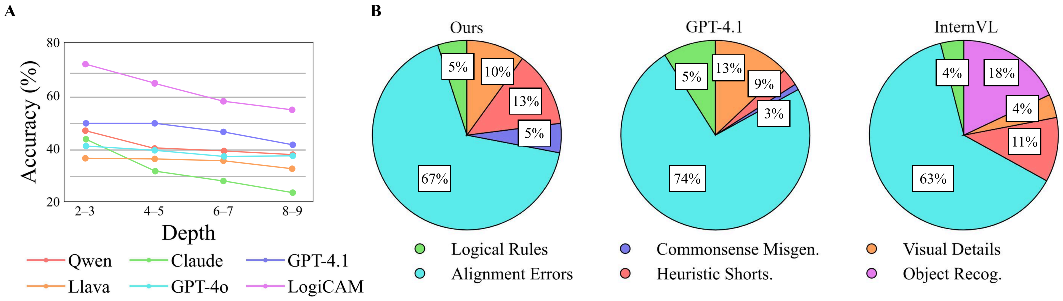


Figure 7: Panel A reports accuracy across different depths, while Panel B illustrates the error distribution across models.

highlights that reasoning-focused objectives improve logical rigor, often at the cost of natural phrasing. This suggests that optimizing for surface or semantic metrics alone isn't enough to improve logical coherence, and future research should target the quality of symbolic logic.

7.3 Depth Analysis

As shown in Figure 7A, **All models exhibit a clear decline in performance as the symbolic reasoning depth increases**, confirming the benchmark's effectiveness in exposing the growing complexity of multimodal logical tasks. GPT-4.1 emerges as the strongest baseline, with the highest accuracy after LogiCAM and a moderate 16% drop from 2–3 to 8–9 steps. However, it still struggles at greater depths, revealing limits in complex multi-hop reasoning. GPT-4o and Llava maintain stable performance with minor 3–4% drops, but their overall accuracy is much lower, indicating a trade-off between robustness and reasoning capacity. In contrast, Claude suffers a sharp 20% decline, highlighting poor generalization on longer symbolic chains.

In contrast, LogiCAM not only delivers superior average performance but also scales more effectively when reasoning chains grow. It demonstrates the strongest overall performance and robustness, consistently outperforming other models across all reasoning depths. It achieves 71.91% accuracy at the shallowest level and maintains a solid 54.61% even at the deepest. Notably, it surpasses the strongest baseline GPT-4.1 by 13% at depths 8–9, highlighting a substantial advantage in handling extended reasoning chains. While LogiCAM exhibits a larger absolute drop across depths, its high performance at all levels indicates strong generalization to both moderate and complex symbolic reasoning tasks. This drop, however, suggests there is still room to improve long-chain reasoning robustness.

7.4 Ablation Study

We conduct an ablation study, which demonstrates that each module is indispensable, as shown in Figure 8A. Removing the symbolic reasoning module produces the largest performance reduction (5.14%), underscoring the importance of adhering to formal logical rules. Omitting heuristic reasoning yields a 3.45% degradation, indicating that heuristics serve as an effective complement when strict logical rules are inapplicable. Disabling premise selection results in a 3.27% drop, reflecting its crucial role in identifying critical information and simplifying subsequent inference. Collectively, these findings highlight that each module plays a critical and non-redundant role, underscoring the necessity of the full design for achieving strong overall performance.

7.5 Error Analysis

We conduct a thorough error analysis by randomly selecting a domain- and symbol-balanced subset of 100 examples for each model. We identify six major error types: incorrect application of logical rules, failure to supplement with heuristic commonsense knowledge, overlooking critical visual details, logical misalignment between visual and textual context, improper reliance on heuristic shortcuts where symbolic reasoning is required, and misperception of objects in the image. Details of each error type are discussed in Appendix A.

Error distribution across different models. As shown in Figure 7 B, failures to logically align and integrate visual with textual premises overwhelmingly dominate (67% for LogiCAM, 74% for GPT-4.1, and 63% for InternVL), demonstrating that cross-modal grounding remains the principal hurdle. Looking specifically at each model:

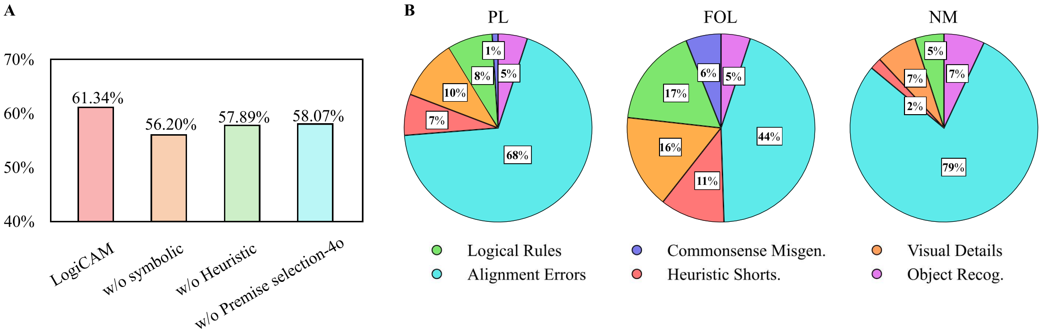


Figure 8: Panel A shows the ablation study results as bar plots, while Panel B presents pie charts illustrating the error distribution across different logical types.

- **LogiCAM** is designed to blend symbolic deduction with heuristic inference; it exhibits a high rate of heuristic shortcuts (13%), indicating difficulty in discerning when to apply formal logic versus commonsense reasoning.
- **GPT-4.1** shows minimal reliance on heuristics (3%) and almost no failures to supplement with commonsense (1%), yet overlooks visual details in 13% of cases and misapplies formal logical rules 9% of the time. The latter aligns with known Chain-of-Thought behavior, where outputs can seem plausible but contain subtle logical errors [37].
- **InternVL** suffers the highest proportion of pure perception errors (18%), reflecting weaker object recognition than GPT-4.1, and relies on heuristic shortcuts in 11% of cases.

Notably, all models suffer major logical misalignment between modalities and visual oversight errors, underscoring a critical need for advances in vision–language fusion. Future work should focus on improving cross-modal fusion and incorporating logic-based training objectives, enabling more accurate symbolic reasoning across modalities.

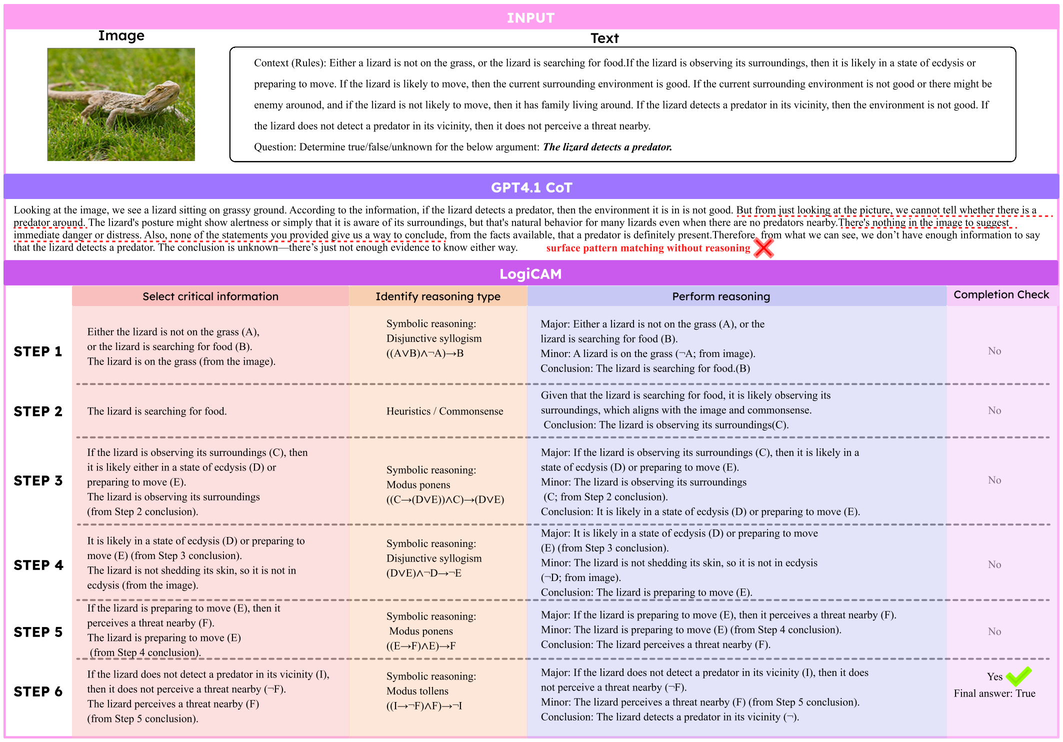
Error distribution across different logical types. We further analyze the error by logical types as shown in Figure 8B, and have the following findings:

- **Consistent Alignment Issues Across Logic Types.** A primary source of failure in PL, FOL, and NM arises from logical misalignment between text and image, with this problem being particularly severe in NM (79%) and PL (68%). This aligns with our broader finding that mapping formal logical structures onto multimodal contexts remains a fundamental challenge for current vision-language models (VLMs).
- **FOL is Most Prone to Overlooking and Logical Errors.** Overlooking errors are most frequent in FOL (16%), where models often miss details in multi-entity, nested, or quantified reasoning. Logical rule errors are also highest (17%), reflecting the symbolic complexity of quantifier binding, variable tracking, and relational reasoning compared to PL or NM.
- **PL's Dependence on Symbolic Alignment.** Although PL avoids many deep logical errors, its performance is highly dependent on accurate logical text-image alignment, as reflected in the 68% rate of alignment errors. Once alignment is achieved, the relatively simple structure of PL facilitates more reliable rule application by the models.
- **NM's High Alignment Difficulty but Low Logical Error Rates.** Despite exhibiting the highest rate of alignment errors (79%), NM shows the lowest incidence of incorrect logical rule application (5%) and commonsense supplementation errors (0%). This pattern suggests that once alignment is successfully established, NM reasoning is more consistent with the model's intuitive understanding or default interpretive patterns, which may partly explain its comparatively strong raw performance.

7.6 Case Study

To illustrate the limitations of existing VLMs and how LogiCAM addresses them, we present a case study comparing the reasoning of GPT-4.1 (with CoT prompting) and LogiCAM in Figure 9.

GPT-4.1's CoT reasoning exhibits a form of "nearsightedness". As the reasoning chain grows longer, it gradually loses the thread that connects image cues to abstract premises, defaulting instead to surface-level judgments (e.g., "I can't see a predator, so unknown"). Without a systematic **Premise Selection** process, it fails to ground observations like "on grass" in relevant textual logical rules (e.g., (not on the grass (A) or searching for food (B)) $\land$ (on the grass ($\neg$A)) $\rightarrow$



INPUT

Image

Text

Context (Rules): Either a lizard is not on the grass, or the lizard is searching for food.If the lizard is observing its surroundings, then it is likely in a state of ecdysis or preparing to move. If the lizard is likely to move, then the current surrounding environment is good. If the current surrounding environment is not good or there might be enemy arounod, and if the lizard is not likely to move, then it has family living around. If the lizard detects a predator in its vicinity, then the environment is not good. If the lizard does not detect a predator in its vicinity, then it does not perceive a threat nearby.

Question: Determine true/false/unknown for the below argument: ***The lizard detects a predator.***

GPT4.1 CoT

Looking at the image, we see a lizard sitting on grassy ground. According to the information, if the lizard detects a predator, then the environment it is in is not good. But from just looking at the picture, we cannot tell whether there is a predator around. The lizard's posture might show alertness or simply that it is aware of its surroundings, but that's natural behavior for many lizards even when there are no predators nearby.There's nothing in the image to suggest immediate danger or distress. Also, none of the statements you provided give us a way to conclude, from the facts available, that a predator is definitely present.Therefore, from what we can see, we don't have enough information to say that the lizard detects a predator. The conclusion is unknown—there's just not enough evidence to know either way. **surface pattern matching without reasoning** ✗

LogiCAM

	Select critical information	Identify reasoning type	Perform reasoning	Completion Check
STEP 1	Either the lizard is not on the grass (A), or the lizard is searching for food (B). The lizard is on the grass (from the image).	Symbolic reasoning: Disjunctive syllogism $((A \lor B) \land \neg A) \to B$	Major: Either a lizard is not on the grass (A), or the lizard is searching for food (B). Minor: A lizard is on the grass ($\neg A$; from image). Conclusion: The lizard is searching for food.(B)	No
STEP 2	The lizard is searching for food.	Heuristics / Commonsense	Given that the lizard is searching for food, it is likely observing its surroundings, which aligns with the image and commonsense. Conclusion: The lizard is observing its surroundings(C).	No
STEP 3	If the lizard is observing its surroundings (C), then it is likely either in a state of ecdysis (D) or preparing to move (E). The lizard is observing its surroundings (from Step 2 conclusion).	Symbolic reasoning: Modus ponens $((C \to (D \lor E)) \land C) \to (D \lor E)$	Major: If the lizard is observing its surroundings (C), then it is likely in a state of ecdysis (D) or preparing to move (E). Minor: The lizard is observing its surroundings (C; from Step 2 conclusion). Conclusion: It is likely in a state of ecdysis (D) or preparing to move (E).	No
STEP 4	It is likely in a state of ecdysis (D) or preparing to move (E) (from Step 3 conclusion). The lizard is not shedding its skin, so it is not in ecdysis (from the image).	Symbolic reasoning: Disjunctive syllogism $(D \lor E) \land \neg D \to \neg E$	Major: It is likely in a state of ecdysis (D) or preparing to move (E) (from Step 3 conclusion). Minor: The lizard is not shedding its skin, so it is not in ecdysis ($\neg D$; from image). Conclusion: The lizard is preparing to move (E).	No
STEP 5	If the lizard is preparing to move (E), then it perceives a threat nearby (F). The lizard is preparing to move (E) (from Step 4 conclusion).	Symbolic reasoning: Modus ponens $((E \to F) \land E) \to F$	Major: If the lizard is preparing to move (E), then it perceives a threat nearby (F). Minor: The lizard is preparing to move (E) (from Step 4 conclusion). Conclusion: The lizard perceives a threat nearby (F).	No
STEP 6	If the lizard does not detect a predator in its vicinity (I), then it does not perceive a threat nearby ($\neg F$). The lizard perceives a threat nearby (F) (from Step 5 conclusion).	Symbolic reasoning: Modus tollens $((I \to \neg F) \land F) \to \neg I$	Major: If the lizard does not detect a predator in its vicinity (I), then it does not perceive a threat nearby ($\neg F$). Minor: The lizard perceives a threat nearby (F) (from Step 5 conclusion). Conclusion: The lizard detects a predator in its vicinity ($\neg$).	Yes ✓ Final answer: True

Figure 9: A Case Study Comparing CoT and LogiCAM

searching for food (B). Moreover, lacking step-by-step formal inference, it eventually abandons the deeper reasoning chain altogether, falling back to superficial pattern matching.

In contrast, LogiCAM systematically derives new knowledge and reaches the correct answer by integrating three tightly-coupled mechanisms at every inference step. Its **Premise Selection** module continuously extracts and logically maps image features into textual element (e.g., "on grass" → food search; "no shedding" → ¬ecdysis), demonstrating its advantages in multimodal fusion. The **Reasoning Type Identifier** then selects the appropriate reasoning type, formal logic for structured inferences (e.g., $C \to (D \lor E)$) or heuristics to complement symbolic logic, thereby balancing the rigor of formal deduction with the flexibility to incorporate knowledge beyond the scope of logic. Finally, the **Symbolic Reasoner** rigorously applies formal inference rules (e.g., disjunctive syllogism, modus ponens, modus tollens) to derive each new conclusion in a systematic and reliable way. This disciplined, iterative process ensures robustness in handling long reasoning chains.

8 Conclusion and Future Work

We have pioneered the **Multimodal Symbolic Logical Reasoning (MuSLR)** task, challenging models to perform precise, rigorous formal logic inferences over combined visual and textual inputs, thereby filling a critical gap in existing benchmarks. To support this research direction, we release **MuSLR-Bench**, a rigorously annotated dataset of 1,093 instances spanning seven application domains, featuring 35 atomic reasoning units and 976 composite logic combinations with depths ranging from 2 to 9. We also propose a strong baseline **LogiCAM**, a novel modular framework that systematically decomposes the reasoning process into premise selection, reasoning-type identification, and formal inference, demonstrating substantial performance gains over prior methods.

Looking forward, our diagnostic analyses reveal two key opportunities for advancing multimodal symbolic reasoning. First, **integrating dedicated symbolic modules is essential**: the LogiCAM outperforms base VLMs precisely because it extracts multimodalities based on logic and embeds explicit symbolic reasoning steps. Second, **existing VLMs struggle to align and fuse visual and textual information when performing formal logic**; Future work should explore tighter multimodal integration, such as cross-modal architectures trained with logic-grounded objectives, to bridge this gap. By making MuSLR and its benchmark publicly available, we hope to catalyze research on these challenges and bring truly rigorous, multimodal symbolic reasoning within reach.

Acknowledgments and Disclosure of Funding

This work is supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001).

References

[1] Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8947–8956, 2019.

[2] Anthropic. Claude 3.7 sonnet. `https://www.anthropic.com/claude/sonnet`, 2025. Accessed: 2025-05-14.

[3] 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv*, 2502:13923, 2025.

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

[5] Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, et al. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the International Conference on Machine Learning*, 2025.

[6] Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Roscoe: A suite of metrics for scoring step-by-step reasoning. *CoRR*, abs/2212.07919, 2022.

[7] Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq R. Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. FOLIO: natural language reasoning with first-order logic. *CoRR*, abs/2209.00840, 2022.

[8] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR)*, volume 2015, pages 991–995, 2015.

[9] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In *Proceedings of the ACL*, pages 1049–1065, 2023.

[10] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Sci Data*, 3:160035, 2016.

[11] Shashank Kirtania, Priyanshu Gupta, and Arjun Radhakrishna. LOGIC-LM++: Multi-step refinement for symbolic formulations. In *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024)*, pages 56–63, Bangkok, Thailand, 2024.

[12] Haoming Li, Zhaoliang Chen, Jonathan Zhang, and Fei Liu. LASP: surveying the state-of-the-art in large language model-assisted AI planning. *CoRR*, abs/2409.01806, 2024.

[13] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013/`.

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

[15] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744, 2023.

[16] Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah Smith, Yejin Choi, and Hannaneh Hajishirzi. Vera: A general-purpose plausibility estimation model for commonsense statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1287, December 2023.

[17] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[18] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *Proceedings of the International Conference on Learning Representations*, 2024.

[19] Theo X. Olausson, Alex Gu, Ben Lipkin, Cedegao E. Zhang, Armando Solar-Lezama, Joshua B. Tenenbaum, and Roger Levy. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, 2023.

[20] OpenAI. Gpt-4o system card. *CoRR*, abs/2410.21276, 2024.

[21] OpenAI. Gpt-4.1. `https://openai.com/index/gpt-4-1/`, 2025. Accessed: 2025-05-14.

[22] Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3806–3824, 2023.

[23] Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 13679–13707, 2024.

[24] Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. Multi-LogiEval: Towards evaluating multi-step logical reasoning ability of large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 20856–20879, 2024.

[25] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649, 2015.

[26] Hyun Ryu, Gyeongman Kim, Hyemin S. Lee, and Eunho Yang. Divide and translate: Compositional first-order logic translation and verification for complex logical reasoning. In *Proceedings of the International Conference on Learning Representations*, pages 236–265, 2025.

[27] Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *Proceedings of the International Conference on Learning Representations*, 2023.

[28] Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3621–3634, 2021.

[29] Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. In *Proceedings of the International Conference on Learning Representations*, 2024.

[30] Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Cheng Jiayang, Chunkit Chan, and Yangqiu Song. CANDLE: Iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2351–2374, 2024.

[31] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.

[32] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 24824–24837, 2022.

[33] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal LLM. In *Proceedings of the International Conference on Machine Learning*, pages 53366–53397, 2024.

[34] Xiaoqian Wu, Yong-Lu Li, Jianhua Sun, and Cewu Lu. Symbol-llm: Leverage language models for symbolic system in visual human activity reasoning. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2023.

[35] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal LLM logical reasoning benchmark in visual contexts. *CoRR*, abs/2407.04973, 2024.

[36] Jundong Xu, Hao Fei, Meng Luo, Qian Liu, Liangming Pan, William Yang Wang, Preslav Nakov, Mong-Li Lee, and Wynne Hsu. Aristotle: Mastering logical reasoning with A logic-complete decompose-search-resolve framework. *CoRR*, abs/2412.16953, 2024.

[37] Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 13326–13365, 2024.

[38] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *CoRR*, abs/2504.15279, 2025.

[39] 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*, pages 9556–9567, 2024.

[40] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SkeHuCVFDr`.

[41] Ruiwen Zhou, Wenyue Hua, Liangming Pan, Sitao Cheng, Xiaobao Wu, En Yu, and William Yang Wang. Rulearena: A benchmark for rule-guided reasoning with llms in real-world scenarios. *CoRR*, abs/2412.08972, 2024.

[42] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, Dengnian Chen, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479, 2025.

Appendix

In the appendix, we provide detailed descriptions of each error type (Section A), the complete workflow of the MuSLR construction pipeline (Section B), the full quality control process, including both automatic and manual filtering strategies (Section C), the details of the LogiCAM framework (Section D), the collection of atomic symbolic logic used in our study (Section F), and an ethics statement (Section G).

A Error Analysis

We provide detailed explanations of each error type below.

Incorrect Application of Logical Rules This error occurs when the model attempts to apply formal logical rules but does so incorrectly. Typical mistakes include reversing implications, confusing necessary and sufficient conditions, or failing to properly follow multi-step deductions. While the model recognizes that logical reasoning is needed, the specific application is flawed, leading to invalid conclusions.

Failure to Supplement with Commonsense / Rule Misgeneralization In some cases, the given input lacks complete information, requiring the model to draw on commonsense knowledge to fill in gaps. This error happens when the model fails to do so, resulting in halted or incomplete reasoning. Alternatively, the model may overgeneralize a formal rule, applying it too broadly or narrowly, which also leads to incorrect outcomes.

Overlooking Visual Details This error reflects the model's inability to notice or correctly interpret critical visual elements in the image, such as small objects, specific colors, or spatial relationships. Missing these details prevents the model from correctly progressing in its reasoning chain, despite the necessary information being present in the visual input.

Premise Integration / Alignment Errors Even when the model successfully extracts information from both text and image, it sometimes fails to align them correctly. This happens when visual entities are mismatched with their textual references (e.g., linking "the red triangle" to the wrong object in the image). Such misalignment breaks the reasoning process and leads to incorrect answers.

Heuristic Shortcuts over Formal Logic Rather than following precise logical reasoning, the model occasionally defaults to heuristic-based shortcuts, relying on superficial patterns or associations learned during training. While this may sometimes produce plausible answers, it undermines the rigor required for formal logical tasks, resulting in systematic errors when heuristics are misapplied.

Visual Perception / Object Recognition Errors This error type stems from failures in basic visual perception, such as misidentifying objects, misclassifying shapes, colors, or spatial positions. When the model starts reasoning from an incorrect visual premise, all subsequent deductions are built on a faulty foundation, leading to incorrect conclusions.

B MuSLR Construction Process

We collect images from multiple sources, including COCO [14], Flickr30k [25], nocaps [1], Mimic [10], RVL_CDIP [8], ScienceQA [17] and Traffic Report collected manually. For each image I, visual details V are extracted using GPT-4o to ensure diverse and fine-grained descriptions.

Step 1: Systematic Rule Selection

We begin by examining a broad set of logical inference rules drawn from propositional logic (PL), first-order logic (FOL), and non-monotonic logic (NM). We utilize the complete set of logical rules collected by [24], denoted as $\mathcal{R} = \{r_1, r_2, \ldots, r_m\}$, which comprehensively covers standard inference patterns. Rather than selecting rules randomly, we carefully curate a subset $\mathcal{R}_{\text{selected}} \subseteq \mathcal{R}$ that is both formally sound and frequently encountered in real-world reasoning. This subset includes classical patterns such as Modus Ponens, Hypothetical Syllogism, Modus Tollens, and Disjunctive Syllogism. Details about the logical rules are provided in the Appendix.

Step 2: Meaningful Rule Composition:

We select meaningful rule combinations, denoted as $\mathcal{R}_{\text{set}} = \{R_1, R_2, \ldots\}$, to construct logically coherent reasoning chains $\mathcal{C} = \{C_1, C_2, \ldots\}$ by rule-based substitution. Each reasoning chain C_i

consists of an ordered sequence of rules from $\mathcal{R}_{\text{set}}$ and is manually composed by experts in formal logical reasoning to ensure coherence and meaningfulness.

Step 3: Grounding in Real-World Contexts:

The meaningful rule composition step produces an abstract, context-independent symbolic rule set $\mathcal{R}_{\text{set}} = \{R_1, R_2, \ldots\}$ (e.g., "If A, then B"). During grounding, visual features V from an image I guide the retrieval of relevant textual information $T_{\text{retrieved}}$ from sources like healthcare reports, Wikipedia, or traffic incident summaries. Abstract rules from $\mathcal{R}_{\text{set}}$ are instantiated using real-world information from $T_{\text{retrieved}}$, creating the grounded rule set $\mathcal{R}_{\text{real}}$ (e.g., "If someone is blowing out candles, they might be celebrating a birthday").

The adapted rule set $\mathcal{R}_{\text{real}}$ will be used to construct the instantiated reasoning chain $\mathcal{C}_{\text{real}}$. When the symbolic reasoning rule $\mathcal{R}_{\text{real}}$ alone is insufficient to capture the real-world context $T_{\text{retrieved}}$, we incorporate commonsense reasoning to supplement formal logic. This combination forms a hybrid reasoning structure $\mathcal{C}_{\text{hybrid}} = (r_1, r_2, \ldots, r_k)$, where each $r_i \in \mathcal{R}_{\text{sym}} \cup \mathcal{R}_{\text{cs}}$. Here, $\mathcal{R}_{\text{sym}}$ comprises rules instantiated from $\mathcal{R}_{\text{set}}$, and $\mathcal{R}_{\text{cs}}$ denotes commonsense reasoning steps. Commonsense reasoning is incorporated only in $\mathcal{C}_{\text{hybrid}}$ and not explicitly represented in $\mathcal{R}_{\text{real}}$. This reflects human cognitive processes, where not all necessary information is always available, and intuitive reasoning is often used to fill in the gaps. The $\mathcal{R}_{\text{real}}$ populates the hybrid reasoning template $\mathcal{C}_{\text{hybrid}}$, yielding the fully grounded reasoning chain $\mathcal{C}_{\text{real}}$. Then we use the conclusion of the $\mathcal{C}_{\text{real}}$ to construct questions and ground-truth answers based on rule-based substitution.

Step 4: Question Generation

Based on the ground-truth reasoning chain C_{gt} and answer A_{gt}, we generate corresponding questions Q that require multi-step reasoning for solution, following rule-based substitution templates.

Step 5: Automatic and Manual Quality Verification

Finally, both automatic verification procedures and manual expert review are employed to ensure the overall quality, consistency, and correctness of the generated dataset.

C MuSLR Quality Check

To ensure the high quality, relevance, and correctness of the constructed dataset, we implement a multi-layered quality control procedure combining both automatic and manual verification steps.

Automatic Quality Control: We apply two automatic filtering strategies to enforce logical soundness and diversity:

- **Lexical Similarity Filtering:** We compute the lexical similarity between each pair of reasoning steps within a reasoning chain using Jaccard Similarity. Chains with a similarity score above 0.5 are discarded to promote step diversity and minimize redundancy.
- **Commonsense Plausibility Filtering:** Each reasoning step is assessed using Vera [16], a T5 model fine-tuned on commonsense reasoning tasks. If any step receives a plausibility score below 0.5, the entire instance is removed to ensure logical soundness and realism.

As a result of the automatic filtering, the original sample size was reduced from 1,956 to 1,464.

Manual Quality Control: Given that the extraction of visual details (V) leverages GPT-4o, which may have hallucinations, we implement a rigorous manual validation stage:

- **Visual Detail Verification:** Human annotators confirm that the extracted visual details accurately reflect the content of the corresponding image, explicitly checking for hallucinated objects, actions, or attributes.
- **Context and Question Evaluation:** Annotators evaluate whether the generated context (T_{context}) and associated questions (Q) are plausible and relevant to real-world scenarios.

Annotation Process and Training All instances were independently reviewed by three trained annotators with STEM backgrounds. In total, six annotators were recruited to assess the 1,464 instances, with each annotator reviewing 732 instances. For each check, annotators provided judgments using a three-option scale: `Yes`, `No`, or `Not Sure`.

To prepare annotators and ensure consistent application of quality standards, we provided a dedicated training session. This session covered task definitions, annotation guidelines, and hands-on practice with feedback. To further support annotators and minimize cognitive load, we developed a custom annotation interface prototype (see Figure 10), which streamlined the annotation process by integrating image previews, visual details, and context input fields for both checks. This tool helped reduce annotation errors and improve task efficiency.

Annotators also underwent a calibration phase involving 30 examples, followed by iterative discussion sessions to refine annotation guidelines and resolve disagreements. We measured inter-annotator agreement using **Fleiss' Kappa**, achieving an average score of 0.92 for visual detail verification (substantial agreement) and 0.71 for context alignment (moderate agreement), which is consistent with the subjective complexity of evaluating real-world plausibility.

Annotation Results. Visual detail verification exhibited a high level of agreement, with an initial inter-annotator agreement rate of approximately 0.90, reflecting the objective nature of the task. In contrast, context alignment showed lower agreement, with an initial rate of around 0.70, due to its inherently more subjective nature. Instances were initially retained if they received three `Yes` votes for both checks.

Conflict Resolution and Filtering.

- Instances that received unanimous `No` judgments from all annotators in either check were directly discarded.
- For cases with conflicting judgments (e.g., one `No`, two `Yes` or any instance with at least one `Not Sure`), a second round of annotation was conducted. During this phase, annotators collaboratively revisited the flagged cases, discussed discrepancies, and reached a consensus decision to ensure consistent quality standards.
- If, after discussion, the final decision still resulted in a `No` for either the visual detail correctness or context plausibility, the instance was removed.

Filtering Statistics and Error Examples: Across the dataset, 492 instances were filtered by automatic checks, 371 by manual annotation, resulting in the final sample size of 1093. Common errors detected included hallucinated objects or implausible contexts, further emphasizing the necessity of both automated and human oversight to ensure dataset validity.

D Detailed LogiCAM Reasoning Process

Below, we present the step-by-step reasoning workflow of LogiCAM.

Step 1: Initial Premise Selection. Given a context set $\mathcal{R}_{\text{real}}$, an image I, and access to a VLM, we prompt the model to initiate the reasoning process by selecting relevant information $I_{\text{relevant}} \subseteq \mathcal{C} \cup \mathcal{V}$. The VLM is instructed to prioritize selecting a pair (ϕ, ψ) such that a formal inference rule (e.g., Modus Ponens) can be applied. If no such pair exists, the model selects the information it judges most critical for solving the task.

Step 2: Identify Reasoning Type. For each selected pair I_{relevant}, we determine the type of reasoning. Symbolic reasoning is applied if the I_{relevant} contain a pair (ϕ, ψ) such that a formal inference rule (e.g., Modus Ponens) can be applied, i.e., $\phi \wedge (\phi \rightarrow \chi) \vdash \chi$. Otherwise, commonsense reasoning is used.

Step 3: Perform Reasoning. Depending on the reasoning type identified in the previous step, the VLM performs inference to derive new knowledge K. For symbolic reasoning, the system applies *syllogistic inference*, a form of deductive reasoning. Specifically, given two selected premises $I_{\text{relevant}} = \{\phi, \psi\}$, the VLM applies formal logical rules to derive a conclusion. For commonsense reasoning, the VLM generates a semantically and contextually plausible implication χ, such that $(I_{\text{relevant}} \rightarrow \chi)$, grounded in real-world commonsense knowledge using a VLM. The result of either reasoning process is recorded as K.

Step 4: Check for Completion. We evaluate whether the current knowledge K is sufficient to determine an answer to the given question. For truth evaluation (True/False/Unknown) questions involving a single hypothesis H, if $K \models H$ or $K \models \neg H$, the process terminates with the corresponding label (`True` or `False`); otherwise, it continues. For multiple-choice questions with candidate hypotheses $\{H_1, H_2, H_3, H_4\}$, we apply the reasoning process to each H_i individually and select the one for which $K \models H_i$ holds, if exactly one such H_i exists. If no hypothesis is entailed, or more than one is, we continue the reasoning process. In all cases, the set of relevant information is updated as $I_{\text{relevant}} \leftarrow I_{\text{relevant}} \cup K$, and the procedure is repeated from Step 1. The reasoning loop is bounded by a predefined number of maximum iterations. If no conclusive answer is reached within this limit, the final output is labeled as `Unknown` for truth evaluation questions, or deemed incorrect for multiple-choice questions.

E Additional Experiments

E.1 Using Symbolic Prover on MuSLR

Most existing LLM+solver approaches (e.g., Logic-LM, Logic-LM++, LINC) are designed for text-only reasoning tasks and cannot directly process visual inputs. Extending them to multimodal settings typically requires a vision-language model (VLM), such as GPT-4.1, to translate images into textual descriptions. However, this translation often omits subtle or hard-to-verbalize visual cues.

To illustrate this limitation, we adapted a representative LLM+solver method, Logic-LM [22], by pairing it with a VLM (GPT-4.1) to convert images into text, and compared its performance against LogiCAM on propositional logic (PL) and first-order logic (FOL). (Logic-LM does not support natural language with modalities, NM.) The results are summarized below:

Model	PL (%)	FOL (%)
Logic-LM + VLM	35.14	32.65
LogiCAM	60.44	42.55

Table 2: Performance comparison of Logic-LM with VLM versus LogiCAM on MuSLR.

These findings demonstrate that simply translating visual information into text is insufficient for effective symbolic reasoning. LogiCAM, which is natively built on VLMs, achieves significantly higher performance since it can directly access visual content. Nonetheless, LLM+solver approaches remain important, and we propose exploring more integrated multimodal LLM+solver frameworks as promising directions for future work.

F Atomic Symbolic Logic

Below, we present the atomic symbolic rules used to construct MuSLR.

Propositional and First-order Logic

- **Modus Ponens (MP)**
  Propositional:
  $$\big((p \to q) \land p\big) \;\vdash\; q$$
  First-order:
  $$\big((\forall x\,(p(x) \to q(x))) \land p(a)\big) \;\vdash\; q(a)$$
  If “p implies q” and p holds, we may conclude q.
- **Modus Tollens (MT)**
  Propositional:
  $$\big((p \to q) \land \neg q\big) \;\vdash\; \neg p$$
  First-order:
  $$\big((\forall x\,(p(x) \to q(x))) \land \neg q(a)\big) \;\vdash\; \neg p(a)$$
  From $p \to q$ and $\neg q$ infer $\neg p$.
- **Hypothetical Syllogism (HS)**
  Propositional:
  $$\big((p \to q) \land (q \to r)\big) \;\vdash\; (p \to r)$$
  First-order:
  $$\big(\forall x\,((p(x) \to q(x)) \land (q(x) \to r(x)))\big) \;\vdash\; (p(a) \to r(a))$$
  Chaining two implications into one.
- **Disjunctive Syllogism (DS)**
  Propositional:
  $$\big((p \lor q) \land \neg p\big) \;\vdash\; q$$
  First-order:
  $$\big((\forall x\,(p(x) \lor q(x))) \land \neg p(a)\big) \;\vdash\; q(a)$$
  Eliminate a disjunct once the other is shown false.

- **Constructive Dilemma (CD)**
  Propositional:
  $$\big((p \to q) \land (r \to s) \land (p \lor r)\big) \ \vdash \ (q \lor s)$$
  First-order:
  $$\big((\forall x\,((p(x) \to q(x)) \land (r(x) \to s(x))) \land (p(a) \lor r(a)))\big) \ \vdash \ (q(a) \lor s(a))$$
  From two conditionals and a choice of antecedents, infer a choice of consequents.
- **Destructive Dilemma (DD)** Propositional:
  $$\big((p \to q) \land (r \to s) \land (\neg q \lor \neg s)\big) \ \vdash \ (\neg p \lor \neg r)$$
  First-order:
  $$\big((\forall x\,((p(x) \to q(x)) \land (r(x) \to s(x))) \land (\neg q(a) \lor \neg s(a)))\big) \ \vdash \ (\neg p(a) \lor \neg r(a))$$
  The “dual” of the constructive dilemma.
- **Biconditional Dilemma (BD)**
  Propositional:
  $$\big((p \to q) \land (r \to s) \land (p \lor \neg s)\big) \ \vdash \ (q \lor \neg r)$$
  First-order:
  $$\big((\forall x\,((p(x) \to q(x)) \land (r(x) \to s(x))) \land (p(a) \lor \neg s(a)))\big) \ \vdash \ (q(a) \lor \neg r(a))$$
  A mix of constructive and destructive patterns.
- **Commutativity of $\lor$ (CT)**
  Propositional:
  $$(p \lor q) \ \dashv\vdash \ (q \lor p)$$
  First-order:
  $$\forall x\,(p(x) \lor q(x)) \ \dashv\vdash \ \forall x\,(q(x) \lor p(x))$$
  Order of a disjunction doesn’t matter.
- **De Morgan’s Transformation (DMT)**
  Propositional:
  $$\neg(p \land q) \ \dashv\vdash \ (\neg p \lor \neg q)$$
  First-order:
  $$\neg\forall x\,(p(x) \land q(x)) \ \dashv\vdash \ \exists x\,(\neg p(x) \lor \neg q(x))$$
  Pushing negation inside a conjunction (or quantifier).
- **Conjunction of Conclusions (CO)**
  Propositional:
  $$\big((p \to q) \land (p \to r)\big) \ \vdash \ \big(p \to (q \land r)\big)$$
  First-order:
  $$\forall x\big((p(x) \to q(x)) \land (p(x) \to r(x))\big) \ \vdash \ \forall x\,(p(x) \to (q(x) \land r(x)))$$
  From two implications with the same antecedent, fuse their consequents.
- **Implication Conjunction (IM)**
  Propositional:
  $$(p \to (q \to r)) \ \dashv\vdash \ ((p \land q) \to r)$$
  First-order:
  $$\forall x\,(p(x) \to (q(x) \to r(x))) \ \dashv\vdash \ \forall x\,((p(x) \land q(x)) \to r(x))$$
  Currying/un-currying of implication.
- **Material Implication (MI)**
  Propositional:
  $$(p \to q) \ \dashv\vdash \ (\neg p \lor q)$$
  (No direct first-order analogue listed.)
- **Existential Generalization (EG)** First-order only:
  $$p(a) \ \vdash \ \exists x\, p(x)$$
  From a particular instance infer an existential claim.
- **Universal Instantiation (UI)** First-order only:
  $$\forall x\, p(x) \ \vdash \ p(a)$$
  From a universally quantified claim infer it for an arbitrary constant.

Extended Multi-variable FOL Rules

- **MV1**

$$\forall x \forall y\, ((p(x) \land q(x)) \to r(x,y)) \;\land\; \exists u \exists v\, (p(u) \land \neg r(u,v)) \;\;\vdash\;\; \exists y\, \neg q(y)$$

  If every $p \land q$ yields r, but there is an instance of p where r fails, then that instance must lack q.

- **MV2**

$$\forall x \forall y\, ((p(x) \land q(x)) \to \neg s(x,y)) \;\land\; \forall z\, (r(z) \to p(z)) \;\land\; r(a) \land s(a,b) \;\;\vdash\;\; \neg q(b)$$

  Combines two universally quantified conditionals and a counter-example to force $\neg q(b)$.

- **MV3**

$$\forall x\, \exists y\, \big(p(x) \to q(x,y)\big) \land \forall u \forall v\, \big((q(u,v) \land r(u,v)) \to s(v)\big) \land \exists z \exists k\, (p(z) \land r(z,k)) \;\;\vdash\;\; \exists w\, s(w)$$

  Chaining an existential-conditional, a universal rule, and an example to derive an existential.

- **MV4**

$$\forall x \forall y \forall z\, (p(x,y,z) \to (q(x,z) \lor r(y))) \;\land\; \exists u \exists v \exists w\, (p(u,v,w) \land \neg q(u,w)) \;\;\vdash\;\; \exists s\, r(s)$$

  If p always gives q or r, and for some triple p holds but q fails, then some r must hold.

- **MV5**

$$\forall x\, (p(x) \to \exists y\, r(y,x)) \;\land\; p(a) \;\;\vdash\;\; \exists z\, r(z,a)$$

  From a universal "p implies an r" and one example of p, infer the corresponding existential.

- **MV6**

$$\forall x \forall y\, (p(x,y) \lor q(x,y)) \;\land\; \exists u \exists v\, \neg q(u,v) \;\;\vdash\;\; \exists z \exists w\, p(z,w)$$

  A quantified disjunction plus a counter-example to one disjunct forces the other.

- **MV7**

$$\forall x \forall y\, (p(x,y) \to (q(x) \land r(y))) \;\land\; p(a,b) \;\;\vdash\;\; q(a) \land r(b)$$

  From a universal conditional that yields a conjunction, plus an instance, you get both conjuncts.

Non-monotonic Default-Reasoning Patterns

- **DRS** (Default Reasoning with Several Defaults) Manages cases where multiple default rules apply at once and may conflict, by finding a consistent combination.
- **DRI** (Default Reasoning with Irrelevant Information) Ensures that adding facts unrelated to a default does not block that default's usual conclusion.
- **DRD** (Default Reasoning with a Disabled Default) Shows how the presence of an exception can "turn off" a default that would otherwise fire.
- **DRO** (Default Reasoning in an Open Domain) Adapts defaults to settings where not all individuals are known or named.
- **REI** (Reasoning about Unknown Expectations I) Allows inferring a default property in the absence of any information to the contrary.
- **REII** (Reasoning about Unknown Expectations II) Refines REI by handling the situation where conflicting expectations might arise.
- **REIII** (Reasoning about Unknown Expectations III) Extends the previous patterns to nested or higher-order expectations.
- **RAP** (Reasoning about Priorities) Introduces a priority ordering among defaults to resolve conflicts in favor of the higher-priority rule.

G Ethics Statement

G.1 Statement

This study adheres to a rigorous ethical framework to ensure the responsible development, evaluation, and deployment of multimodal general-purpose AI models. The key ethical considerations are outlined below. These measures ensure that MuSLR, as a responsible and inclusive framework, continuously contributes to the fair, sustainable, and accountable development of multimodal artificial intelligence.

G.2 Privacy and Data Protection

The benchmarking and evaluation processes strictly comply with privacy regulations. All tasks and datasets used in MuSLR are carefully curated to exclude any personally identifiable information (PII). Any data obtained from publicly available sources is anonymized and filtered to remove privacy-sensitive content. We are committed to fully adhering to relevant data protection standards, including the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), thereby upholding the highest standards of ethical research practices.

G.3 Data Collection

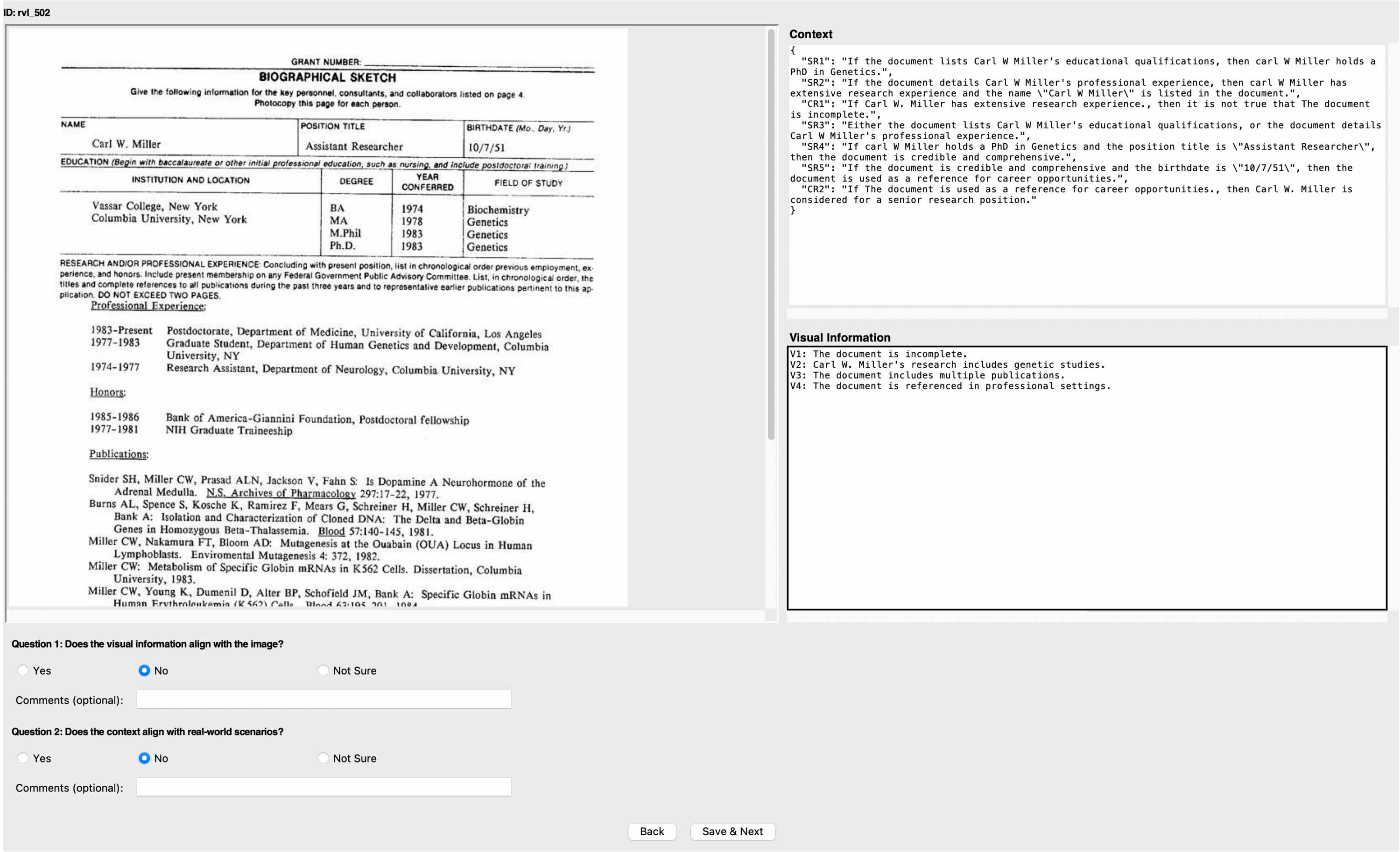
All data included in the MuSLR dataset was sourced exclusively from publicly available resources. The data collection protocol is designed to prioritize ethical sourcing, ensuring that contributors' rights are respected, including the right to withdraw their data where applicable. This approach ensures transparency and fairness throughout the dataset construction process.

G.4 Annotator Compensation

We fully recognize the critical role human annotators play in creating the high-quality MuSLR dataset. All six annotators involved in the project are trained professionals, and they received fair compensation for their work. Annotators were compensated with cash payments upon completion of their assigned tasks. Each annotator was committed to contributing their best efforts to data annotation and quality assurance, ensuring the integrity and reliability of the dataset.

G.5 Bias and Fairness

We proactively implemented measures to analyze and mitigate potential biases related to gender, ethnicity, language, and other sociocultural factors present in the datasets and evaluation tasks. Our goal is to reduce the risk of perpetuating biases in AI development. While completely eliminating bias remains an ongoing challenge, our commitment to identifying and addressing bias throughout the benchmark development process remains steadfast.



ID: rvl_502

GRANT NUMBER: ______

BIOGRAPHICAL SKETCH

Give the following information for the key personnel, consultants, and collaborators listed on page 4.
Photocopy this page for each person.

NAME	POSITION TITLE	BIRTHDATE (Mo., Day, Yr.)
Carl W. Miller	Assistant Researcher	10/7/51

EDUCATION (*Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.*)

INSTITUTION AND LOCATION	DEGREE	YEAR CONFERRED	FIELD OF STUDY
Vassar College, New York	BA	1974	Biochemistry
Columbia University, New York	MA	1978	Genetics
	M.Phil	1983	Genetics
	Ph.D.	1983	Genetics

RESEARCH AND/OR PROFESSIONAL EXPERIENCE: Concluding with present position, list in chronological order previous employment, experience, and honors. Include present membership on any Federal Government Public Advisory Committee. List, in chronological order, the titles and complete references to all publications during the past three years and to representative earlier publications pertinent to this application. DO NOT EXCEED TWO PAGES.

Professional Experience:

1983-Present	Postdoctorate, Department of Medicine, University of California, Los Angeles
1977-1983	Graduate Student, Department of Human Genetics and Development, Columbia University, NY
1974-1977	Research Assistant, Department of Neurology, Columbia University, NY

Honors:

1985-1986	Bank of America-Giannini Foundation, Postdoctoral fellowship
1977-1981	NIH Graduate Traineeship

Publications:

Snider SH, Miller CW, Prasad ALN, Jackson V, Fahn S: Is Dopamine A Neurohormone of the Adrenal Medulla. N.S. Archives of Pharmacology 297:17-22, 1977.

Burns AL, Spence S, Kosche K, Ramirez F, Mears G, Schreiner H, Miller CW, Schreiner H, Bank A: Isolation and Characterization of Cloned DNA: The Delta and Beta-Globin Genes in Homozygous Beta-Thalassemia. Blood 57:140-145, 1981.

Miller CW, Nakamura FT, Bloom AD: Mutagenesis at the Ouabain (OUA) Locus in Human Lymphoblasts. Enviromental Mutagenesis 4: 372, 1982.

Miller CW: Metabolism of Specific Globin mRNAs in K562 Cells. Dissertation, Columbia University, 1983.

Miller CW, Young K, Dumenil D, Alter BP, Schofield JM, Bank A: Specific Globin mRNAs in Human Erythroleukemia (K562) Cells. Blood 63:195-201, 1984.

Context

```
{
  "SR1": "If the document lists Carl W Miller's educational qualifications, then carl W Miller holds a PhD in Genetics.",
  "SR2": "If the document details Carl W Miller's professional experience, then carl W Miller has extensive research experience and the name \"Carl W Miller\" is listed in the document.",
  "CR1": "If Carl W. Miller has extensive research experience., then it is not true that The document is incomplete.",
  "SR3": "Either the document lists Carl W Miller's educational qualifications, or the document details Carl W Miller's professional experience.",
  "SR4": "If carl W Miller holds a PhD in Genetics and the position title is \"Assistant Researcher\", then the document is credible and comprehensive.",
  "SR5": "If the document is credible and comprehensive and the birthdate is \"10/7/51\", then the document is used as a reference for career opportunities.",
  "CR2": "If The document is used as a reference for career opportunities., then Carl W. Miller is considered for a senior research position."
}
```

Visual Information

```
V1: The document is incomplete.
V2: Carl W. Miller's research includes genetic studies.
V3: The document includes multiple publications.
V4: The document is referenced in professional settings.
```

Question 1: Does the visual information align with the image?

○ Yes ● No ○ Not Sure

Comments (optional):

Question 2: Does the context align with real-world scenarios?

○ Yes ● No ○ Not Sure

Comments (optional):

Figure 10: Annotation Interface. We developed a custom interface to streamline the annotation process and reduce annotator effort.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We explicitly include the contribution sessions and link that to our main claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide detailed error analysis to discuss the limitations of our proposed method.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not have theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We discussed the data construction and method in the main content, and provided more details in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will open-source all the data and codes.

   Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We include the details in the experiment setting.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our benchmark is large so it will be too expensive.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).

- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include the details in the experiment setting.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We fully conduct the NeurIPS code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Not applicable. The proposed benchmark is a technical contribution intended for research evaluation and does not have direct societal impacts, either positive or negative.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no such risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper and specify the version we used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide sufficient details about the benchmark in Section 4.

    Guidelines:

    - The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We include the details of how we instruct the human annotators in appendix.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We discuss how we minimize human subjects in appendix.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We discuss the details of LLM usage in the paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.