Visual–Linguistic Abductive Reasoning with LLMs for Knowledge-based Visual Question Answering

Jieun Kim[1] **Yujin Jeong**[2] **Sung-Bae Cho**[2]
[1]Dept. of Artificial Intelligence and [2]Dept. of Computer Science, Yonsei University
{lilly9928,yujinj00,sbcho}@yonsei.ac.kr

Abstract

Recent attempts to leverage large language models (LLMs) for reasoning and pre-trained knowledge in multi-modal reasoning focus on two main approaches: aligning image features with linguistic space, and converting images into textual cues to exploit the implicit reasoning capabilities of LLMs. Although they integrate visual information into the reasoning pipeline, they often treat visual perception and language reasoning as separate processes, limiting the potential for fully unified multi-modal reasoning. In this paper, we propose a novel method, Visual–Linguistic Abductive Reasoning (ViLA), inspired by human abductive reasoning processes. ViLA hypothesizes a plausible answer, generates the corresponding visual and textual premises, and employs fuzzy scoring to select the most coherent combination, thus deriving the final inference. This process integrates visual and linguistic modalities into interpretable abductive reasoning chains, enabling unified multi-modal reasoning. Without fine-tuning LLMs or retrieving external knowledge, ViLA improves performance by 2.31% on AOKVQA, 1.7% on OKVQA, and 1.7% on GQA over previous state-of-the-art models, while also improving interpretability and stability.

1 Introduction

When humans solve complex visual problems, they often rely on abductive reasoning (Peirce, 1936) to infer the most plausible explanation from incomplete or uncertain observations. For example, when we encounter an image-related question as shown in Figure 1, we typically perform the following steps to solve the problem. First, we examine the image and notice that a person is wearing ski equipment, leading us to hypothesize that the sport is skiing. Second, we generate possible answer candidates, such as snow and ski, that could plausibly explain both the visual scene and the question.

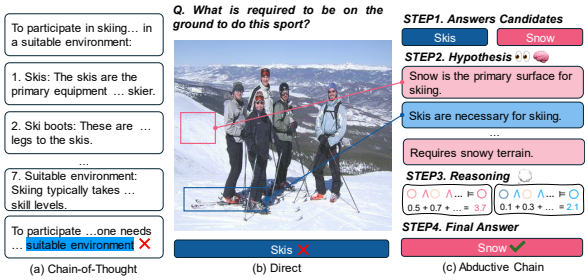


Figure 1: An example to introduce our motivation. Unlike (a) chain-of-thought or (b) direct prediction, (c) abductive reasoning integrates visual cues and contextual knowledge to generate and evaluate plausible hypotheses, leading to more accurate integrated visual–textual reasoning.

Third, we combine these candidates with contextual knowledge to assess which explanation best fits the evidence. Fourth, we select the candidate with the highest plausibility, resulting in the final answer. This reflects how humans perform visual abductive reasoning by generating hypotheses from visual cues, justifying them with contextual knowledge, and selecting the best explanation under uncertainty.

Recent advances in large language models (LLMs) have demonstrated that chain-of-thought (CoT) prompting enables step-by-step reasoning that approximates human cognitive processes (Wei et al., 2022). Extending such reasoning to multi-modal scenarios requires models to integrate information from both visual and textual modalities in a unified manner. To employ the reasoning capabilities of LLMs for multi-modal tasks, two primary approaches have been proposed: Aligning visual features with linguistic representations to enable joint reasoning across modalities (Liu et al., 2023; Li et al., 2023; Ye et al., 2024; Lin et al., 2024), and converting visual information into textual descriptions to facilitate in-context learning within language models (Cho et al., 2021; Yang

Findings of the Association for Computational Linguistics: EACL 2026, pages 6529–6544

et al., 2022; Chen et al., 2024b). Although these approaches have proven effective in handling visual–textual input, visual reasoning often involves incomplete or uncertain observations, making it difficult to manage the inherent uncertainty in such scenarios. Furthermore, achieving seamless integration of reasoning across text, images and implicit knowledge remains challenging, often resulting in missing question intent, under-utilization of relevance knowledge, or omission of critical visual information (see Figure 1 (a),(b)).

To address these challenges, we propose Visual–Linguistic Abductive Reasoning (ViLA), a method inspired by human abductive reasoning, which models hypothesis generation and evaluation across modalities rather than performing formal logical abduction. ViLA first hypothesizes plausible answers using visual cues and contextual information, then constructs supporting and competing premises from both modalities to ground these hypotheses, and finally applies abductive reasoning with fuzzy plausibility scoring to handle uncertainty and select the most coherent interpretation of the visual–linguistic input.

ViLA enhances interpretability by explicitly approximating the abductive reasoning process, enables integrated multi-modal reasoning, and improves robustness via fuzzy logic-based evaluation of competing hypotheses. In summary, this paper makes the following contributions.

- **Novelty:** We introduce ViLA, a novel method for human-like visual abductive reasoning that generates and evaluates hypotheses under uncertainty.
- **Integrated multi-modal reasoning:** ViLA connects visual perception and language reasoning through logical rule representations and abductive inference.
- **Efficient computation:** It preserves modality-specific information without full-text conversion and leverages VLM and LLM complementarily, ensuring efficiency without additional training or performance loss.

2 Related Works

In this section, we explore the relevant work in three key areas: abductive reasoning, reasoning in a visual language model (VLM), and knowledge-based visual question answering (VQA).

2.1 Abductive Reasoning

Abductive reasoning, introduced by Peirce as inference to the best explanation, is fundamental to narrative comprehension and causal inference. In NLP, (Hobbs et al., 1993) formalized language understanding as weighted abduction, while (Zhao et al., 2023) introduced the first commonsense-driven benchmark. Subsequent work explored counterfactual reasoning (Qin et al., 2020; Rudinger et al., 2020), missing fact generation (Tafjord et al., 2021), annotation-free learning, and knowledge graph correction (Bai et al., 2024). In contrast, multimodal abductive reasoning is less explored: Early studies proposed visual benchmarks (Hessel et al., 2022; Liang et al., 2022), with recent extensions to atypical events (Chinchure et al., 2025) and image entailment (Ventura et al., 2025).

2.2 Reasoning in Vision-Language Models

Large pre-trained models have enabled multimodal reasoning (Lu et al., 2019; Tan and Bansal, 2019; Marino et al., 2021; Wu et al., 2022; Kim et al., 2025). Recent work extends chain-of-thought reasoning to multimodal contexts (Zhang et al., 2023), either by aligning image and language spaces (Liu et al., 2023; Li et al., 2023; Ye et al., 2024; Lin et al., 2024; Alayrac et al., 2022) or converting images into text for direct LLM input (Yang et al., 2022; Hu et al., 2023; Shao et al., 2023). However, these approaches face challenges such as catastrophic forgetting, hallucination, and loss of dense visual information (Zheng et al., 2023; Chung and Yu, 2023; Chen et al., 2020), which limits their ability to support interactive reasoning.

2.3 Knowledge-based Visual Question Answering

Knowledge-based VQA (Marino et al., 2019; Schwenk et al., 2022) requires combining image understanding with external knowledge. Early methods retrieved from structured knowledge bases, while recent approaches leverage the implicit knowledge of LLMs to improve performance (Yang et al., 2022; Lin et al., 2022; Shao et al., 2023; Chen et al., 2024b). In this paper, we build on this line of research and propose a model that integrates multiple modalities to facilitate interactive reasoning. To validate our method, we focus on the knowledge-based VQA problem.

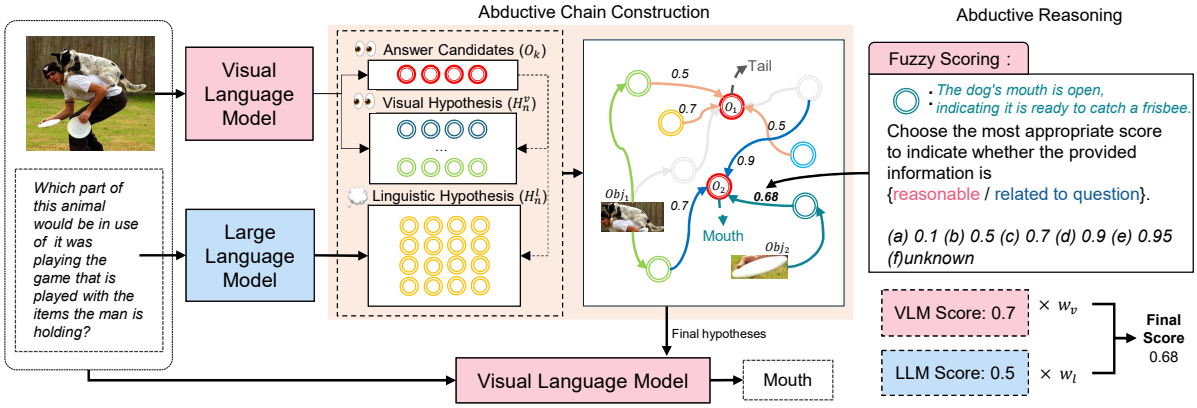


Figure 2: Overview of the proposed method. Given an image–question pair, the model performs abductive reasoning by hypothesizing plausible answers, generating visual and linguistic premises, and selecting the most coherent explanation through fuzzy scoring for unified multi-modal reasoning.

3 The Proposed Method

ViLA approximates human abductive reasoning by modeling how plausible explanations are generated and evaluated across modalities. As illustrated in Figure 2 and Algorithm 1, the system first generates candidate answers and their corresponding visual premises using a VLM, then refines the linguistic premises through an LLM. Each premise is independently evaluated via fuzzy scoring within the two models, and their scores are integrated to estimate overall validity. Premises exceeding a predefined threshold are retained as final hypotheses to derive the conclusion. Detailed procedures for each step are described in the following sections.

3.1 Task Formulation

This paper aims to evaluate a model that approximates human reasoning behavior by solving a knowledge-based Visual Question Answering (VQA) task requiring multimodal reasoning across visual and textual information. Conventional Vision–Language Models (VLMs), as formulated in Equation (1), address this task by directly predicting the answer Y given a question X_Q and an image X_V:

$$\hat{Y} \sim P_\theta(\cdot \mid X_Q, X_V) \tag{1}$$

Although this formulation focuses on maximizing the accuracy of the answer, it provides limited interpretability of the reasoning process behind the prediction.

In contrast, we redefine the task to explicitly incorporate reasoning evidence into the prediction process. Specifically, the model infers the final answer based not only on the question and image but also on the derived reasoning premises H_{final}, which serve as structured justifications for the answer:

$$Y \sim P_\theta(\cdot \mid X_Q, X_V, H_{\text{final}}) \tag{2}$$

This reformulation shifts the goal from merely generating answers to reasoning-based prediction, allowing explicit evaluation of the model's inferential behavior. The detailed procedure for generating and evaluating H_{final} is described in the following sections.

3.2 Abductive Chain Construction

We employ B-graphs (Gallo et al., 1993; Wu et al., 2024) to construct abductive chains. A B-graph is a directed hypergraph composed of B-arcs, each connecting multiple premises to a single conclusion. Formally, it is defined as (S, R), where S is a set of symbols and R a set of rules $r = (P, c)$ with premises $P \subseteq S$ and conclusion $c \in S$, expressed as $r := \bigwedge_{p \in P} \models c$. In our method, B-graphs serve as the structural backbone for entailment-based chain generation. The premise–conclusion pairs (P, c) are generated by entailment-aware prompting, allowing the abductive chain to expand dynamically within this formal representation.

3.2.1 Answer Candidate Generation

The process of generating conclusions in ViLA follows the general methodology of conventional VQA, where a VLM predicts answers based on the provided image and question. Unlike standard approaches, however, ViLA does not treat the generated answer as the final answer. Instead, it uses them as a set of preliminary answer candidates. To examine whether the VLM beam output contains

Algorithm 1 Pipeline of the proposed ViLA

Input: Image and Question $\{I, Q\}$
Output: Answer $\hat{Y}$

$P_v, P_l, P_{final} \leftarrow \emptyset$
$n \leftarrow \#premises$
$C_k \leftarrow ConclusionGenerator(I, Q)$
$i \leftarrow 0$
# Logic Rule Generation
while $i < n$ **do**
  $v_i \leftarrow VLMPremises(I, Q, C_k)$
  $l_i \leftarrow LLMPremises(Q, C_k)$
  $P_v \leftarrow P_v \cup \{v_i\}$
  $P_l \leftarrow P_l \cup \{l_i\}$
  $i \leftarrow i + 1$
end while
$P \leftarrow P_v \cup P_l$
$PremiseLength \leftarrow len(P)$
$j \leftarrow 0$
# Logical Reasoning
while $j < PremiseLength$ **do**
  $\mu_R(P_j) \leftarrow f(I; P_j)$
  $\mu_A(P_j) \leftarrow f(Q; P_j)$
  $\mu_V(P_j) \leftarrow f_{Con}(\mu_R(P_j), \mu_A(P_j))$
  if $\mu_F(P_j) \geq t$ **then**
    $P_{final} \leftarrow P_{final} \cup \{P_j\}$
  end if
  $j \leftarrow j + 1$
end while
$\hat{Y} \leftarrow VLMPredict(I, Q, P_{final})$
return $\hat{Y}$

the correct answer, we performed experiments using ten beams. Figure 3 illustrates both the overall accuracy in varying beam sizes and the proportion of correct answers captured within each beam. The results reveal that, while the correct answer frequently appears in the beam candidates, it is often discarded during the final selection stage, hindering the accurate answer. To address this issue, ViLA directly selects answer candidates from the beam output, ensuring that potentially correct answers are retained for subsequent abductive reasoning. The specific generation method is presented in Equation (3). The set O of candidate answers comprises k answers o, which are selected by applying beam search to the VLM model in order to identify the top k most probable answers.

$$O_k = \text{argmax}_{O \subseteq Y, |O|=k} \prod_{i=1}^{N} P_\theta(Y_i \mid X_Q, X_V) \tag{3}$$

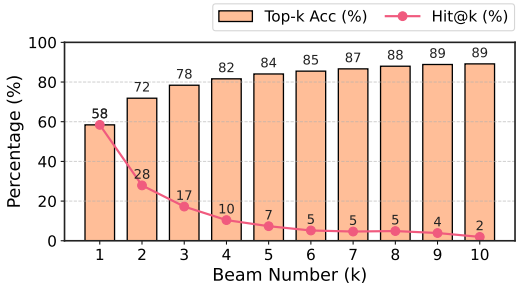


Figure 3: Top-k accuracy and Hit@k for different beam sizes. The table evaluates answer coverage and ranking behavior under varying beam sizes.

3.2.2 Premise Construction

The premise generation process is carried out based on the generated candidate answers, using both VLM and LLM. Initially, the VLM uses the given image X_V and the generated candidate answers o_k to extract evidence from the image that could substantiate the answer, with this evidence serving as the premise H_i^o.

$$H^v = P_{VLM}(h_i^v | X_V, o_k) \tag{4}$$

Subsequently, the LLM generates textual premises p_i^t related to the given answer o_k, leveraging its explicit knowledge to produce relevant justifications.

$$H^l = P_{LLM}(h_i^l | X_Q, o_k) \tag{5}$$

The premises generated by the VLM and LLM are then combined to form the final set of premises H, which is used to fuzzy-score the final premises.

$$h^v, h^l \in H \tag{6}$$

Through this premise generation process, we aim to explicitly harness the inherent reasoning capabilities of the model and make the reasoning process transparent by grounding it in logical rules.

3.3 Fuzzy Scoring for Abductive Reasoning

In the process of abductive reasoning, fuzzy scoring is applied to evaluate the plausibility of the generated premises. The fuzzy sets used in this process consist of three components: (1) *adequacy*, which represents the correspondence between the image and the premise; (2) *relevance*, which represents the correspondence between the question and the premise; and (3) *validity*, which integrates adequacy and relevance to represent the overall soundness of the hypothesis. These are defined as follows:

$$\mu_{\text{R}}(H), \mu_{\text{A}}(H) \in \{0.1, 0.5, 0.7, 0.9, 0.95\},$$
$$\mu_{\text{V}}(H) \in [0, 1].$$

The discrete membership values correspond to linguistic levels of fuzzy confidence: *very low* (0.1), *low* (0.5), *medium* (0.7), *high* (0.9), and *very high* (0.95). Here, the intermediate value 0.5 is assigned to the 'low' level rather than 'medium' to reflect a conservative interpretation of the responses of uncertain models. Empirically, both VLMs and LLMs tend to produce moderate confidence values even when the entailment is weak; thus, mapping 0.5 to 'low' helps prevent overestimation and yields more stable fuzzy reasoning.

In ViLA, membership degrees are computed using both a vision–language model (VLM) and a language model (LLM) as follows:

$$\begin{aligned} \mu_A(H) &= f_{\text{vlm}}(X_V, H), \\ \mu_R(H) &= f_{\text{llm}}(X_Q, H). \end{aligned} \tag{7}$$

The degree of membership is obtained through a prompt-based scoring process defined as:

Fuzzy Scoring

"Choose the most appropriate score to indicate whether the information provided is {variable}. Return only the letter corresponding to your choice from the options below:
(a) 0.1 (b) 0.5 (c) 0.7 (d) 0.9 (e) 0.95 (f) unknown"

Here, the *variable* depends on the model: for VLM, it is 'reasonable to the image,' and for the LLM, it is 'relevant to the question.' To estimate a unified membership value from two fuzzy sets, a convex combination is used, a common method in fuzzy reasoning for aggregating multiple degrees of membership into a single representative value. Consequently, the fuzzy set *validity* integrates adequacy and relevance as follows:

$$\mu_{\text{V}}(H) = w_{\text{v}} \cdot \mu_{\text{A}}(H) + w_{\text{l}} \cdot \mu_{\text{R}}(H), \quad w_{\text{v}} + w_{\text{l}} = 1, \tag{8}$$

where w_{v} and w_{l} represent the relative weights assigned to adequacy and relevance, respectively. This formulation enables the validity score to capture the combined effect of visual and linguistic consistency while preserving the normalized range of fuzzy membership values.

Finally, hypothesis H is accepted as the final premise H_{final} if its validity exceeds a threshold t:

$$H_{final} = \begin{cases} P, & \mu_V(H) \geq t, \\ \neg P, & \text{otherwise.} \end{cases} \tag{9}$$

4 Experiments

4.1 Implementation Details

The proposed method utilizes off-the-shelf VLM and LLM models, incorporating in-context learning techniques. For the experimental setup, unless otherwise specified, all results employ the LLaVA-NeXT (Mistral7B) model as the VLM and the Mistral7B model as the LLM. In all tables and figures, 7B refers to Mistral7B, 13B refers to LLaVA-NeXT (13B), and GPT4 refers to gpt4o-mini. To enhance object recognition within images, a pre-trained fast R-CNN model is utilized during the initial phase of object extraction. The number of premises and conclusions generated by ViLA is treated as a hyperparameter, both set to 2 for the experiments, and the threshold for selecting the final premise is set to 0.7. For w_v and w_l is set to 0.9 and 0.1. Furthermore, to facilitate step-by-step reasoning, the widely adopted CoT method is applied to both the base model, LLaVA, and the proposed ViLA for evaluation. All experiments were conducted on a cluster of five NVIDIA A100.

4.1.1 Baselines

We compare ViLA with representative vision-language models and reasoning-based methods:

Visual Language Models: BLIP2, InstructBLIP, and LLaVA-NeXT are baseline models that align visual features with LLMs for visual question answering.

Chain-of-Thought (CoT) applies a reasoning strategy where image information extracted from each baseline is converted into text and processed through an LLM using CoT prompting for fair comparison.

SoTA Methods: ViperGPT, VisProg, and VCTP are state-of-the-art methods that perform interactive inference by combining VLMs and LLMs.

4.1.2 Datasets and Evaluation Metrics

We evaluate ViLA on three VQA benchmarks that require external knowledge and complex reasoning, following the evaluation protocol for each dataset. Datasets dependent on explicit knowledge graphs are excluded, as ViLA focuses on interactive reasoning using commonsense knowledge combined with visual and linguistic modalities.

OKVQA (Marino et al., 2019) is a widely used knowledge-based VQA dataset containing 14,055 image-text pairs. The questions require reasoning

Methods	OK-VQA	A-OKVQA
Direct Answering		
MAVEx (Wu et al., 2022)	41.37	-
UnifER (Guo et al., 2022)	42.13	-
ViLBERT (Lu et al., 2019)	35.20	30.60
LXMERT (Tan and Bansal, 2019)	36.91	30.70
KRISP (Marino et al., 2021)	38.40	33.70
PICa (Yang et al., 2022)	48.00	43.37
Prophet (Shao et al., 2023)	61.10	58.20
LION (Chen et al., 2024a)	57.33	60.87
QACap (Yang et al., 2025)	68.20	66.30
Structured Reasoning		
ViperGPT (Surís et al., 2023)	51.90	39.50
VisProg* (Gupta and Kembhavi, 2023)	41.84	52.12
VCTP (Chen et al., 2024b)	56.20	53.20
Base VLMs		
BLIP2 (Li et al., 2023)	32.32	38.94
InstructBLIP (Dai et al., 2023)	61.16	62.23
LLaVA-NeXT (Liu et al., 2023)	55.61	66.48
+ CoT (7B)	13.34	34.10
+ CoT (GPT-4)	30.56	52.71
Base VLMs + ViLA		
BLIP2 + **ViLA (7B)**	34.05	39.37
InstructBLIP + **ViLA (7B)**	61.51	62.31
LLaVA-NeXT + **ViLA (7B)**	57.92	68.18
LLaVA-NeXT-13B + **ViLA (GPT-4)**	**68.91**	**69.98**

Table 1: Comparison of accuracy (%) on OK-VQA and A-OKVQA datasets.

Methods	Knowledge Sources	GQA
LXMERT	LXMERT	60.0
BLIP2	Vicuna-13B	41
InstructionBLIP	Vicuna-7B	49.2
InstructionBLIP	Vicuna-13B	49.5
ViperGPT	GPT-3	48.1
VisProg	CLIP, GPT-3	50.5
LLaVA-NeXT	Mistral-7B	58.74
+ CoT	Mistral-7B	32.1
+ **ViLA**	Mistral-7B	**63.59**

Table 2: Comparison of accuracy (%) on GQA dataset.

that combines visual content with external knowledge such as common sense.

A-OKVQA (Schwenk et al., 2022) is the largest knowledge-based VQA dataset, providing questions paired with rationales. It emphasizes complex reasoning grounded in both visual understanding and external knowledge.

GQA (Hudson and Manning, 2019) is used to further analyze the limitations of ViLA. It categorizes questions and requires multi-step interactive reasoning for accurate predictions.

4.2 Quantitative Results

In Table 1, we compare model performance on the OK-VQA and A-OKVQA datasets. Using LLaVA-NeXT as the baseline, ViLA achieves improvements of 2.31% and 1.7%, respectively. As shown in Figure 3, when evaluating up to beam 2, the accuracy reaches 72%, while the proposed method attains 69.98%, demonstrating effective utilization of the vision-language model (VLM). By incorporating an abductive reasoning chain, ViLA enhances cross-modal interaction beyond conventional in-context and zero-shot approaches. Compared to neuro-symbolic methods such as ViperGPT and VisProg, which depend on explicit program generation for reasoning, LLaVA-NeXT-ViLA (7B) achieves gains of +6.02% and +16.08% on OK-VQA, and +28.68% and +16.06% on A-OKVQA, respectively. This indicates that abductive reasoning grounded in logical rules provides a more efficient mechanism for integrating visual and linguistic evidence than program-based pipelines.

However, converting visual information into purely textual representations for Chain-of-Thought (CoT) reasoning—using LLaVA-NeXT for caption generation and GPT-4 for reasoning—leads to significant performance degradation. As shown in Table 1, LLaVA-NeXT + CoT (GPT-4) achieves only 30.56% on OK-VQA and 52.71% on A-OKVQA, demonstrating the loss of visual grounding when multi-modal signals are collapsed into text. In contrast, ViLA preserves modality-specific representations and integrates them through abductive reasoning chains, yielding consistently higher accuracy and interpretability.

Additional evaluations on the GQA test-dev set (Table 2) corroborate these findings: ViLA outperforms the LLaVA-CoT baseline by 4.85% and surpasses in-context and zero-shot models without task-specific fine-tuning. Overall, representing each modality as logical premises within an abductive reasoning chain enhances cross-modal interaction and enables unified, interpretable multi-modal reasoning.

Figure 4 provides a detailed evaluation of ViLA's reasoning behavior. The left graph analyzes the impact of reasoning depth, showing that performance consistently improves over baseline in all steps. This suggests that ViLA's abductive reasoning chains enable effective multi-step inference by integrating visual and linguistic premises. The right graph breaks down performance by question type. Accuracy gains are observed for both binary and open-ended queries, indicating that the abductive reasoning process generalizes across different task formats. Furthermore, the *Distribution* metric (Dist) reflects the chi-square statistic between the predicted and ground-truth answer distributions.

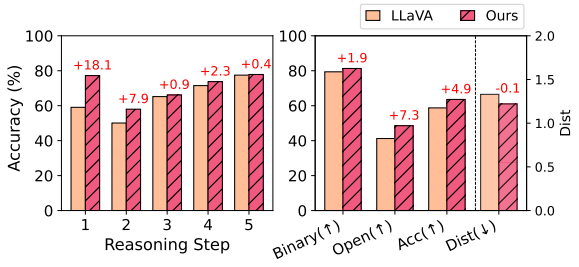


Figure 4: Accuracy by reasoning steps and question types. (Left) Effect of reasoning depth. (Right) Performance across question types.

Methods	LLM	Sentence Similarity
CoT	Mistral7B	38.7
CoT*	Mistral7B	40.9
CoT	GPT4o-mini	39.02
ViLA	Mistral7B	**44.8**

Table 3: Rationale performance comparsion of ViLA and CoT baseline on A-OKVQA validation set.

ViLA achieves a lower chi-square value, suggesting that its predictions are better aligned with the true answer distribution, leading to more balanced and less biased outputs across diverse types of question.

Table 3 presents the results of similarity evaluations on the generated rationales, measured using CLIP sentence similarity. The average cosine similarity was calculated between the CLIP (ViT-B/16) text embeddings of the generated and ground-truth rationales. ViLA shows higher similarity scores than the CoT results produced by the same LLM model and models trained on larger datasets such as GPT-4. The CoT results from the Mistral-7B model show the lowest similarity to the ground-truth rationales. In addition, when rationales are generated first and answers are produced subsequently (marked with an asterisk *), the rationale similarity improves by 2.2% compared to direct answer generation.

4.3 Qualitative Results

Figure 5 presents a qualitative analysis of ViLA's reasoning process compared to baseline models, including LLaVA and CoT. In this example, w_l and w_v are set to 0.1 and 0.9, respectively. For the question "Where are the cats sitting on top of?," ViLA generates multiple candidate premises and evaluates them using the adequacy fuzzy set S_a and the relevance fuzzy set S_r. The constructed premises shown in the figure correspond to those that are ultimately selected after fuzzy evaluation.

	Structural reasoning					
Model	**Cho**	**Comp**	**Log**	**Qry**	**Ver**	**Avg**
LLaVA	84.15	67.23	77.93	71.22	81.35	76.38
+ViLA	**85.56**	**68.25**	**79.87**	**78.55**	**83.75**	**79.20**

Table 4: Structural reasoning category evaluation.

	Semantic reasoning					
Model	**Attr**	**Cat**	**Glob**	**Obj**	**Rel**	**Avg**
LLaVA	68.01	37.95	21.66	84.32	51.53	52.69
+ViLA	**71.33**	**51.72**	**64.33**	**86.89**	**54.97**	**65.85**

Table 5: Semantic reasoning category evaluation.

Among them, one example, "Cats are on keyboard", reflects both strong semantic alignment with the question and solid visual grounding. By selecting the optimal conclusion from these high-scoring premises, ViLA correctly predicts 'keyboard,' while both the LLaVA and CoT baselines incorrectly answer 'desk.' This highlights ViLA's ability to identify and retain only the most valid premises through abductive reasoning.

Table 4 and 5 present the breakdown of the accuracy in the categories of structural and semantic reasoning. Within the structural dimensions (*Choose*, *Compare*, *Logical*, *Query*, and *Verify*), ViLA consistently outperforms baseline LLaVA, with the greatest improvement observed in *Query*, indicating stronger performance in extracting relevance information from the visual context. Gains in *Logical* and *Verify* further suggest that ViLA's abductive reasoning enhances consistency checking and logical inference across visual-linguistic premises.

For the semantic categories (*Attribute*, *Category*, *Global*, *Object*, and *Relation*), ViLA achieves substantial improvements, particularly in *Category* and *Global*. The large gain in *Global* reasoning demonstrates ViLA's ability to integrate scene-level information, while improvements in *Category* and *Relation* reflect better semantic abstraction and understanding of inter-object dependencies. Notably, *Object*-level accuracy also improves, suggesting that ViLA's abductive chain reasoning preserves fine-grained visual grounding while enhancing high-level semantic inference. These results collectively indicate that ViLA strengthens both structural reasoning processes and semantic understanding, highlighting the effectiveness of representing visual-linguistic evidence as abductive premises for multi-step reasoning.

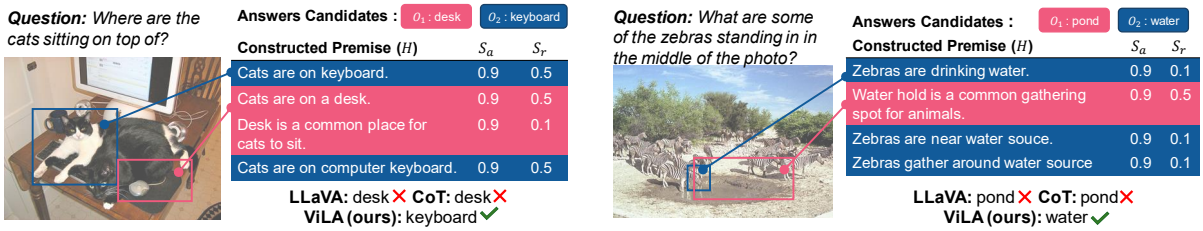


Figure 5: Qualitative results of ViLA's reasoning process. S_a represents the conclusion, while S_r denotes the adequacy fuzzy set score, and S_v corresponds to the relevance fuzzy set score generated through the VLM and LLM.

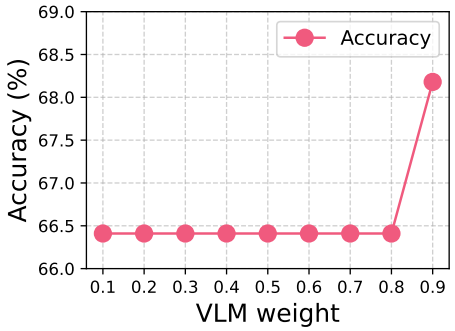


Figure 6: Accuracy variation with VLM weight.

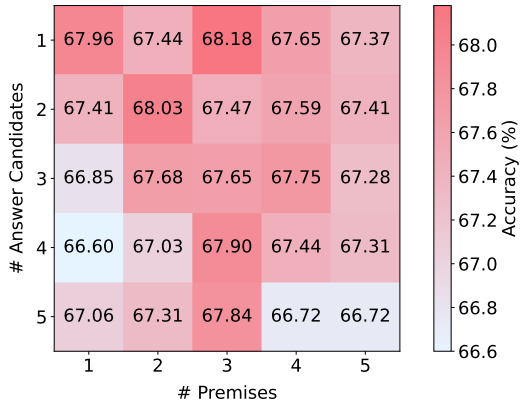


Figure 7: Accuracy by number of conclusions and premises. Color intensity indicates accuracy (%).

LLM premise	VLM premises	Acc(%)
✓		65.33
✓		65.33
✓	✓	**68.18**

Table 6: Ablation study on modality-specific premise generation.

4.4 Ablation Study

As shown in Figure 6, the model maintains a stable accuracy in most weight configurations but shows a noticeable increase when the VLM weight w_v reaches 0.9. This indicates that assigning a higher weight to the VLM, which emphasizes visual adequacy in the validity computation, leads to improved overall reasoning performance. The result suggests that visual evidence plays a more decisive role than textual relevance in constructing valid abductive premises within the ViLA framework.

As shown in 6, we analyze the contribution of modality-specific premises by conducting premise-removal ablation experiments with premises generated by an LLM and a VLM, respectively. We apply the same fuzzy scoring metric across all premise settings to ensure a fair comparison. Using either modality-specific premise alone achieves 65.33% accuracy, whereas combining both premises improves performance to 68.18%, corresponding to a gain of 2.85 percentage points. This suggests that the two modality-specific premises play complementary roles in reasoning for knowledge-base VQA.

We perform an ablation analysis focusing on the premise and answer candidates' hyperparameters. Figure 7 presents the results of the hyperparameter experiments for the A-OKVQA validation dataset. The results indicate that the highest accuracy is achieved when the premise is set to 3 and the answer candidates are set to 1, or when both the premise and the answer candidates are set to 2. In particular, configurations with a premise value of 3 consistently show higher accuracy. This experiment also reveals that the generation of an excessive number of premises and answer candidates leads to the generation of ambiguous supporting data, ultimately resulting in a decrease in accuracy. Therefore, it can be concluded that the use of the appropriate amount of information, rather than overwhelming the model with excessive data, is crucial to improve accuracy.

5 Concluding Remarks

In this paper, we have proposed a Visual–Linguistic Abductive reasoning called ViLA that integrates visual and textual information through abductive reasoning processes similar to human inference. Starting from observations, ViLA generates modality-specific premises and constructs plausible hypotheses, selecting the most coherent explanation via fuzzy scoring to build interpretable abductive reasoning chains that unify visual and linguistic modalities. Extensive experiments on OKVQA, AOKVQA, and GQA demonstrate that this approach not only achieves state-of-the-art accuracy without external knowledge or fine-tuning, but also enhances interpretability and stability in multi-step reasoning. Although ViLA relies on the underlying capabilities of LLMs and VLMs, our results show that abductive reasoning chains can maximize the potential of each model by tightly integrating their modality-specific strengths into a unified reasoning process. Future work is needed to reduce this dependency and make the method more robust across different model architectures.

6 Limitation

Although ViLA achieves improved performance and interpretability, it still depends heavily on the underlying VLM and LLM, making it sensitive to their biases and perception errors. In addition, the absence of external knowledge retrieval can limit its effectiveness in domain-specific reasoning tasks.

Acknowledgments

This work was supported by IITP grant funded by the Korea government (MSIT) (No. RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University); No. RS-2022-II220113, Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework), LG Electronics Inc, and Air Force Defense Research Sciences Program funded by Air Force Office of Scientific Research.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, and 1 others. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 23716–23736.

Jiaxin Bai, Yicheng Wang, Tianshi Zheng, Yue Guo, Xin Liu, and Yangqiu Song. 2024. Advancing abductive reasoning in knowledge graphs through complex logical hypothesis generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1312–1329.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. 2024a. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26540–26550.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. 2024b. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 1254–1262.

Aditya Chinchure, Sahithya Ravi, Raymond Ng, Vered Shwartz, Boyang Li, and Leonid Sigal. 2025. Black swan: Abductive and defeasible video reasoning in unpredictable events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24201–24210.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1931–1942.

Jiwan Chung and Youngjae Yu. 2023. VLIS: Unimodal language models guide multimodal language generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 700–721.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 49250–49267.

Giorgio Gallo, Giustino Longo, Stefano Pallottino, and Sang Nguyen. 1993. Directed hypergraphs and applications. *Discrete Applied Mathematics*, 42(2-3):177–201.

Yangyang Guo, Liqiang Nie, Yongkang Wong, Yibing Liu, Zhiyong Cheng, and Mohan S. Kankanhalli. 2022. A unified end-to-end retriever-reader framework for knowledge-based VQA. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 2061–2069.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14953–14962.

Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The abduction of Sherlock Holmes: A dataset for visual abductive reasoning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 558–575.

Jerry R Hobbs, Mark E Stickel, Douglas E Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63(1-2):69–142.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. 2023. PromptCap: Prompt-guided image captioning for VQA with GPT-3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2963–2975.

Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709.

Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. 2025. Visual question answering: A survey of methods, datasets, evaluation, and challenges. *ACM Computing Surveys*, 57(10).

Junnan Li, Dongxu Bureau Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 19730–19742.

Chen Liang, Wenguan Wang, Tianfei Zhou, and Yi Yang. 2022. Visual abductive reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15565–15575.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. VILA: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26689–26699.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. REVIVE: Regional visual representation matters in knowledge-based visual question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 10560–10571.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 34892–34916.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 13–23.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. KRISP: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14111–14121.

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

Charles Sanders Peirce. 1936. Collected papers of charles sanders peirce. *Nature*, 138:1037–1037.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 794–805.

Rachel Rudinger, Vered Shwartz, Jena D Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 4661–4675.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 146–162.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14974–14983.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. ViperGPT: Visual inference via Python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11888–11898.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics (ACL)*, pages 3621–3634.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5099–5110.

Mor Ventura, Michael Toker, Nitay Calderon, Zorik Gekhman, Yonatan Bitton, and Roi Reichart. 2025. NL-Eye: Abductive NLI for images. In *International Conference on Learning Representations (ICLR)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, pages 2712–2721.

Xiaoqian Wu, Yong-Lu Li, Jianhua Sun, and Cewu Lu. 2024. Symbol-LLM: Leverage language models for symbolic system in visual human activity reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Zhen Yang, Zhuo Tao, Qi Chen, Liang Li, Yuankai Qi, Anton van den Hengel, and Qingming Huang. 2025. Separation of powers: On segregating knowledge from observation in LLM-enabled knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24753–24762.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, pages 3081–3089.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13040–13051.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander Rush. 2023. Abductive commonsense reasoning exploiting mutually exclusive explanations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14883–14896.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. DDCoT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 5168–5191.

7 Appendix

8 Further Quantitative Evaluation

Table 7 and Table 8 presents a comprehensive comparative analysis between the proposed ViLA model and the baseline LLaVA-mistral7b model.

Reasoning Step	LLaVA	ViLA(ours)	Gap
1	59.07	**77.22**	+18.15
2	50.09	**58.01**	+7.92
3	65.21	**66.13**	+0.92
4	71.50	**73.77**	+2.27
5	77.49	**77.86**	+0.37
6	90.24	90.24	0.00
7	95.00	95.00	0.00
8	66.67	66.67	0.00
9	100.00	100.00	0.00

Table 7: Accuracy by reasoning step on the GQA dataset.

Words Number	LLaVA	ViLA(ours)	Gap
3	12.58	**42.38**	+29.80
4	37.30	**57.14**	+19.84
5	40.08	**50.85**	+10.77
6	53.23	**60.03**	+6.80
7	59.14	**63.09**	+3.95
8	61.86	**65.06**	+3.20
9	65.96	**67.99**	+2.03
10	67.65	**68.86**	+1.21
11	64.59	**65.49**	+0.90
12	**72.41**	71.94	-0.47
13	63.64	**65.15**	+1.51
14	71.30	**73.33**	+2.03
15	73.84	**74.26**	+0.42
16	**76.07**	74.36	-1.71
17	**67.02**	64.36	-2.66
18	**80.26**	78.95	-1.31
19	81.40	**83.72**	+2.32
20	**68.75**	62.50	-6.25
21	84.21	84.21	0.00
22	83.33	83.33	0.00
23	25.00	**50.00**	+25.00
24	100.00	100.00	0.00
25	100.00	100.00	0.00

Table 8: Accuracy by number of words on the GQA dataset.

8.1 Reasoning Step Evaluation

The Reasoning Step category evaluates the models' performance across varying levels of reasoning complexity. The ViLA model shows a clear advantage, particularly in the early reasoning steps (1-5), with the most notable improvement of 18.15%p occurring at the 1 reasoning step. However, as the reasoning steps increase to 6 or more, the performance of both models converges, indicating that ViLA's advantage diminishes in more complex, multi-step reasoning scenarios.

8.2 Words Number Evaluation

The Words Number category analyzes the impact of question length on model performance. The ViLA model exhibits significant accuracy improvements for questions ranging from 3 to 15 words, with the most substantial gain of 29.80%p observed for 3-word questions. This suggests that ViLA is particularly effective in understanding and responding to shorter, more concise queries. However, as the question length exceeds 15 words, the performance gap between the two models narrows, and in some cases, LLaVA performs similarly or slightly better than ViLA.

8.3 Analysis of Prompt Sensitivity

Table 9 presents the experimental results comparing different models and their performance across various types of Chain of Thought (CoT) prompts. The results indicate that for models such as LLaVA-Mistral7B and InstructionBlip, the introduction of CoT prompts or longer instructions leads to a decline in accuracy. Specifically, for LLaVA-Mistral7B, accuracy decreased by 2.14%p with CoT and by 2.67%p with long instructions. Similarly, Llava-Vicuna7B showed a more pronounced decline, with a 3.14%p drop with CoT and a 3.79%p drop with long instructions. InstructionBlip exhibited the most significant reduction, particularly when CoT prompts were used, resulting in a 9.34%p decrease in accuracy.

In contrast, ViLA, demonstrates robustness to these variations in prompting. It not only maintained accuracy when CoT or long instructions were introduced but also showed a slight improvement of 0.37%p. The findings from these experiments underscore the potential negative impact of improperly aligned CoT prompts and instructions on LLM-based models, while also highlighting the effectiveness of our approach in preserving and even enhancing model performance.

Model	CoT Type	Acc(DA)
LLaVA-Mistral7B	w/o CoT	68.62
	w CoT	66.48 (-2.14)
	w Long instruction	65.95 (-2.67)
Llava-Vicuna7B	w/o CoT	66.14
	w CoT	63.00 (-3.14)
	w Long instruction	62.35 (-3.79)
Instructionblip	w/o CoT	54.05
	w CoT	44.71 (-9.34)
	w Long instruction	51.41 (-2.64)
ViLA(Ours)	w/o CoT	68.18
	w CoT	68.55 (+0.37)
	w Long instruction	68.55 (+0.37)

Table 9: Accuracy comparison by prompt for different models

8.4 Beam-Search Evaluation on Additional Models

To ensure that the observed beam-search behavior is not limited to LLaVA-NeXT (7B), we applied the same analysis to both LLaVA-NeXT (7B) and InstructBLIP. Figure 8 shows the Top-k accuracy (bars) and Hit@k (dashed lines) for these models as a function of beam size k. In each case, the probability of including the correct answer increases as the beam size grows.

8.5 Ablation Study on Convex weight w_v

Figure 9 visualizes the ablation study comparing the accuracy for different $\boldsymbol{w}$ values when performing a convex combination. The figure demonstrates that the highest performance is achieved when $w_v = 0.9$ This indicates that setting a higher contribution to the adequacy fuzzy set $\mu_{\mathrm{R}}(P)$ leads to improved performance.

9 Case Analysis

9.1 Success Cases

Figure 10 shows more examples of successful cases. For questions like "Is the color of the wristband different than the hat?", it can be seen that the model generates premises such as "The hat and wristband are the same color." or "Wristband matches hat's color." and assigns a high score of 0.9. It is also observed that the model can generate incorrect premises like "The wristband is white, the hat is red," but in these cases, a score of 0 is assigned during the scoring stage, helping the model produce an accurate final conclusion by verifying the generated premise once more. Through this process, it

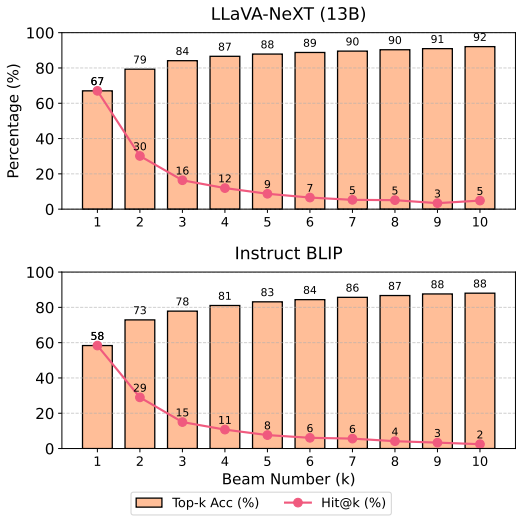


Figure 8: Comparison of Top-k accuracy (bars) and Hit@k (dashed lines) for LLaVA-NeXT (13B) and Instruct BLIP as the beam size k varies.

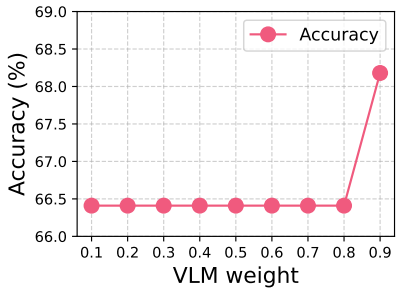


Figure 9: Ablation study on the convex combination weight w_v

is confirmed that the ViLA proposed in this paper improves interpretability in the reasoning process.

9.2 Failure Cases

Figure 11 visualizes a failure case within the GQA dataset, illustrating an instance where the model generated an incorrect response. For the question, "What is lying next to the pens that are lying on top of the bed?" one of the generated premises was "The cell phone is a common item to be found on a bed." This statement exemplifies the model's hallucination, as it introduces false information about an object not present in the image. The associated score of 0.5, while not particularly high, suggests that the model identified some degree of correspondence with objects within the image, leading to the assignment of this score. As depicted in Figure 1, the fuzzy logic-based calculation was somewhat effective in addressing these hallucinations, yet the phenomenon was not entirely eliminated. For the final question, "Is the racket black and long?" the model generated the premises "The racket is red and white." and "The racket is black color." It can be observed that the model assigned a score of 0 to the premise involving "black," indicating a lack of accurate recognition regarding the racket, which is actually a mix of yellow and black. This recognition error appears to have led to the incorrect conclusion of "no" as the final answer.

10 Prompt Templates

10.1 Answer Candidates

- **Input:**`[question]`
- **Template:**

```
Question:[question]
Answer in one word.
```

10.2 Visual Premises

- **Input:** `[num_premises]`, `[start_word]`, `[end_word]`
- **Template:**

```
Provide [num_premises] brief reasons for
    the answer using visual commonsense.
Each explanation must be concise, simple,
    and relevant to the [object].
Use a maximum of 10 words per explanation.
Avoid detailed descriptions and keep the
    explanations straightforward.
Answer in this format: 1. explanation1 2.
    explanation2 ... [num_premises].
    explanation[num_premises].

Question: "[question]"
Answer: "[answer]"

This is because
```

10.3 Linguistic Premises

- **Input:** `[num_premises]`
- **Template:**

```
Provide [num_premises] short explanations
    about the conditions and properties
of the answer and question which can be
    known in image.
Each explanation must be concise, simple,
    and relevant to the image.
Use a maximum of 10 words per explanation.
Avoid detailed descriptions and keep the
    explanations straightforward.

Question: "[question]"
Answer: "[answer]"
```

10.4 Fuzzification ($\mu_{\mathbf{A}}(H)$)

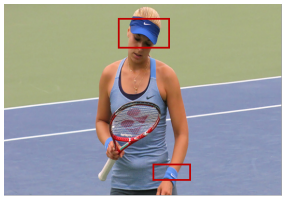
- **Input:** [information], [image]
- **Template:**

```
Given the following:

- Information: "[information]"

Choose the most appropriate score to
    indicate whether the provided
information is reasonable to image.
Return only the letter corresponding to
    your choice from the options
below:

- (a) 0.1
- (b) 0.5
- (c) 0.7
- (d) 0.9
- (e) 0.95
- (f) unknown

Example choice format: (a)
```

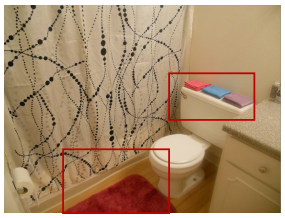
10.5 Fuzzification(($\mu_{\mathbf{R}}(H)$)

- **Input:** [information], [question]
- **Template:**

```
Given the following:

- Information: "[information]"
- Question: "[question]"

Choose the most appropriate score to
    indicate whether provided information
    is related to question.
Return only the letter corresponding to
    your choice from the options
below:

- (a) 0.1
- (b) 0.5
- (c) 0.7
- (d) 0.9
- (e) 0.95
- (f) unknown
```

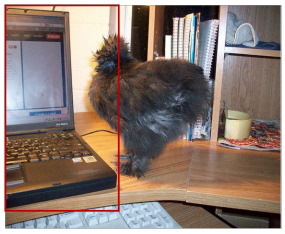


Question: Is the color of the wrist band different than the hat?
Premise: Both are blue, The hat and wristband are the same color, Both are blue, The hat is blue, Both have similar hue, Wristband matches hat's color, The hat is blue, The wristband is white, The hat is blue, The wristband is white, Wristband is blue, Hat is red.
Answer: No



Question: Is the color of the toilet paper different than the color of the mat?
Premise: The toilet paper is pink, The mat is blue, Toilet paper is white, Mat is red, Toilet paper is white, Labels are common on containers, Object is identifiable by shape, The toilet paper is pink, The mat is also pink, The toilet paper is white, The mat is red, Both are white.
Answer: Yes

Question: Is the person that is happy sitting or walking?
Premise: The person is holding a sign, The person is walking, The person is wearing a helmet, which is typically worn while riding a motorcycle, The person is riding a motorcycle, which is a mode of transportation, Smiling face indicates happiness, Person's posture shows movement (walking), The person is holding a sign, The person is sitting, The person is sitting on a motorcycle, The person is smiling, which indicates happiness, The person's facial expression shows happiness, The position of the person suggests they are seated.
Answer: walking



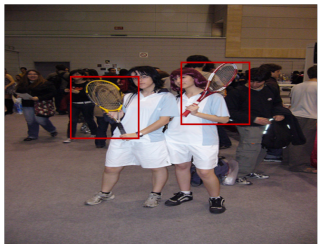
Question: Is that a camera or a laptop?
Premise: The keyboard is visible, The laptop is open, The object is on a desk, The object is a laptop, Has a keyboard, Closed display, Large green rind, Heavier than empty box, Contains bound papers, The keyboard is visible, which is not the case with a laptop, The keys are arranged in a grid pattern, which is typical for a camera, The object is small and black, It is positioned in front of a laptop.
Answer: laptop

Figure 10: Visualization of a success case in GQA.

Question: What is lying next to the pens that are lying on top of the bed?
Premise: The cell phone is a common item found on beds, The pens are also common items found on beds, The cell phone is a common item to be found on a bed, It is a common item to be found next to pens, Electronic device, Commonly found near bedsides, The wallet is a common item to be found on a bed, It is a personal item that is often kept close to pens, which are also commonly found on a bed, The wallet is a common item to be found on a bed, It is a personal item, often kept close to pens for easy access, Bed has linens and pillows, Wallet is a flat object.
Answer: Cell Phone ✖ **GT:** spoon

Question: Are there either juice boxes or mugs in this scene?
Premise: The cups are white and have text on them, which is typical of mugs, The cups are made of paper, which is commonly used for juice boxes, Mugs on table, Juice box in hand, The cups are paper, They are disposable, The image shows a person using a laptop at a table, There are no juice boxes or mugs visible in the scene, Only glasses present, No juice boxes or mugs visible.
Answer: Yes ✖ **GT**: No

Question: What vehicle is on the road?
Premise: The vehicle is large and has a prominent grill, The vehicle is red and white, The image shows a large vehicle on the road, The vehicle is a truck, Large cargo carrier, Heavy duty transportation, The image shows a vehicle on the road, The vehicle is a car, The image shows a red car on the road, The car is a four-door sedan, Four wheels, Common transportation means.
Answer: Truck ✖ **GT:** Car



Question: Is the racket black and long?
Premise: The racket is red and white,The racket is short, The racket is red and white, The racket is short, Racket is white and short, Not suitable for tennis or badminton, The racket is black in color, The racket is long in size, The racket is black in color, The racket is long in size, Racket shape is elongated, Color is black.
Answer: No ✖ **GT:** Yes

Figure 11: Visualization of a failure case in GQA.