Visual Attention Reasoning via Hierarchical Search and Self-Verification

Wei Cai[1,2,*,†], **Jian Zhao**[2,*,‡], **Yuchen Yuan**[2,‡], **Tianle Zhang**[2], **Zheng Zhu**[3], **Haichuan Tang**[3], **Xuelong Li**[2,‡]

[1]Peking University
[2]Institute of Artificial Intelligence (TeleAI), China Telecom
[3]CRRC Academy

Abstract

Multimodal Large Language Models (MLLMs) frequently hallucinate due to their reliance on fragile, linear reasoning and weak visual grounding. We propose Visual Attention Reasoning (VAR), a reinforcement learning framework that reformulates reasoning as a hierarchical search with self-verification. VAR enforces traceable evidence grounding by generating explicit bounding boxes, guided by a novel reward function combining geometric precision and semantic sufficiency. Furthermore, it replaces linear Chain-of-Thought with a tree-search policy capable of backtracking to correct logical errors. Theoretical analysis validates the framework's reliability, and extensive experiments demonstrate that VAR significantly outperforms state-of-the-art methods on complex hallucination and safety benchmarks.

1 Introduction

In recent years, the field of multimodal large language models (MLLMs) has witnessed significant advances (Liu et al., 2023b; Bai et al., 2023; Zhu et al., 2023; Li et al., 2023b; Team et al., 2024; An et al., 2025). Despite their remarkable success, they still suffer from critical limitations that hinder their reasoning capabilities. MLLMs are notoriously prone to visual hallucinations, *i.e.* describing objects or attributes absent in the image (Guan et al., 2024; Liu et al., 2024a, 2023a; Cai et al., 2025b; Jiang et al., 2025). They also heavily rely on linguistic shortcuts, wherein textual priors are favored instead of genuine visual understanding (Si et al., 2022).

More recently, reinforcement learning (RL) methodologies, particularly those inspired by the R1-style framework, have shown promising performances in enhancing the reasoning capabilities of MLLMs across various tasks (Huang et al., 2025; Shen et al., 2025; Zhang et al., 2025). However, these approaches often induce a bias towards "heavy thinking, light observation", an over-reliance on linguistic deliberation at the expense of robust visual perception (Liu et al., 2025; Yao et al., 2025). This imbalance renders MLLMs susceptible to common RL pitfalls such as reward hacking (Fu et al., 2025) and spurious correlations(Shao et al., 2025). Consequently, while RL-trained MLLMs may exhibit ostensible performance gains, they are largely attributed to a superficial distributional shift, where the model's outputs are aligned only with the stylistic training data. This encourages the generation of shortcut answers based on linguistic priors, neglecting the underlying risk of model hallucinations (Li et al., 2025). This issue is seen even on state-of-the-art models, whose performance degrades dramatically when task complexity surpasses a certain threshold (Stechly et al., 2024; Shojaee et al., 2025; Hochlehnert et al., 2025).

Formally, we summarize the issue above as two fundamental limitations. Firstly, the MLLMs lack robust visual grounding; models merely gather superficial features of the image before defaulting to their powerful linguistic priors, resulting in hallucinations or the oversight of critical, nuanced visual details. Secondly, the MLLMs' reasoning process is relatively brittle; a single fallacious step can sabotage the entire linear CoT, leading to a completely invalid conclusion due to the lack of backtracking.

To address these limitations, we begin with the observation that complex reasoning inherently requires a search process within an abstract solution space, where a given reasoning step can seldomly be deterministically derived from its predecessors. Instead, a reasoner typically faces uncertainty and must engage in trial-and-error exploration in several promising directions. This uncertainty is often resolvable only in hindsight that a particular path of inquiry has already been validated as correct or conclusively falsified. Such situations make back-

[*]These authors contributed equally to this work.
[†]Work done during internship at TeleAI.
[‡]Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8986–8997

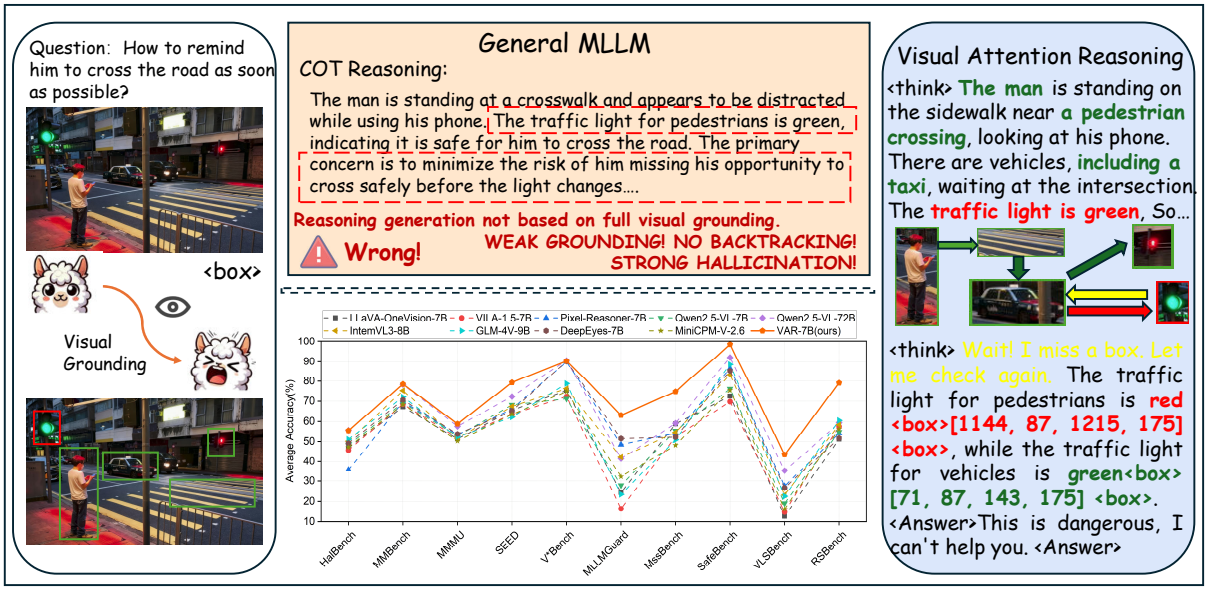

Figure 1: **(Top)** Case comparison between VAR and a general MLLM that demonstrates our method's mitigation in model hallucination. **(Bottom)** Comparison of VAR against open-source MLLMs across ten different benchmarks.

tracking to a prior reasoning juncture and selecting an alternative path especially favorable.

In practice, human experts often implement reasoning backtracking by constructing a structured mental representation of the overall process, which is similar to a reasoning search tree, and navigate it through selective and efficient exploration to avoid combinatorial explosion. Enlightened by this, we introduce visual attention reasoning (VAR), a framework that recasts grounded reasoning not as a linear process, but as a structured searching process over a "reasoning trajectory space." The core idea of VAR can be described as the model's deliberate allocation of cognitive effort, allowing it to explore different reasoning paths, validate intermediate steps, and backtrack from errors. This enables a more robust, multi-step deliberative reasoning process. The main contributions are as follows:

The Visual Attention Reasoning (VAR) Framework: We formalize and implement VAR, a novel framework that decomposes reasoning into traceable evidence grounding and search-based Chain-of-Thought. Its integrated backtracking mechanism directly addresses the brittleness of conventional linear CoT methods.

A Multi-Faceted Self-Verification Reward for Guidance: The search process within VAR is guided by a novel four-component reward function featuring semantic (R_{sem}) and geometric (R_{geo}) self-verifications. This function acts as an internal critic, steering the search away from hallucinatory paths and towards conclusions that are both semantically sufficient and geometrically precise.

Theoretical Guarantees for VAR's Efficiency: A comprehensive analysis is conducted, which proves that the VAR search process is able to find a correct reasoning trajectory with high probability while maintaining a polynomially bounded search space, guaranteeing its controlled efficiency and preventing unbounded computational cost.

A Unified Framework for Traceable Visual Reasoning: We formalize and implement VAR, which to our knowledge is the first framework to unify (1) traceable evidence grounding, (2) a learnable, search-based Chain-of-Thought with backtracking, and (3) a multi-faceted self-verification reward mechanism into a single, cohesive reinforcement learning framework. This integrated design directly addresses the brittleness of conventional linear CoT methods, providing a novel and robust solution to combat MLLM hallucination in complex reasoning tasks.

2 Visual Attention Reasoning

As previously discussed, incorporating intermediate visual supervision is critical for enhancing the reasoning capabilities of MLLMs. However, a practical dilemma is faced by existing approaches: external human annotations are static and expensive, while internal signals lack grounding in visual reality. To address this issue, we introduce a novel RL framework that synergistically combines the

strengths of both paradigms.

Our core idea is to decompose the monolithic visual reasoning process into two distinct, verifiable stages: **1) Traceable Evidence Grounding**, where the model identifies and localizes salient visual evidences, and **2) Search-Based Chain-of-Thought with Backtracking**, where the model reasons over this evidence, backtracking from errors to self-correct.

2.1 The Decomposed Reasoning Trajectory

For a vision-language task $Q = \{i, q\}$, our framework decomposes the reasoning trajectory, denoted as s, into a structured sequence with the following key stages: `<visual_perception>` c `</visual_perception>` `<think>` t `</think>` `<answer>` a `</answer>`

Here, c is the self-contained visual perception, a textual description that must capture all visual information necessary to solve the task, including explicit bounding box coordinates $\{\hat{b}_i\}$ for all relevant objects. t is the subsequent language reasoning trace, and a is the final answer.

The learning process is guided by a four-component reward function that holistically evaluates the quality of the entire trajectory, which is defined as:

$$\begin{aligned} r(Q, s) =& R_{\text{acc}}(a, a^*) + \alpha R_{\text{fmt}}(s) + \\ & \beta R_{\text{sem}}(Q, c) + \gamma R_{\text{geo}}(c, b^*) \end{aligned} \tag{1}$$

where α, β, γ are weighting hyperparameters. Each of the four reward components is explained below:

Accuracy Reward (R_{acc}) is the primary task-level reward, which is defined as $R_{\text{acc}}(a, a^*) = \mathbb{I}[a = a^*]$, where a^* is the ground-truth answer. It provides the ultimate supervisory signal for the entire reasoning process.

Format Reward (R_{fmt}) is a standard binary reward that penalizes trajectories deviating from the required syntactic structure.

Semantic Verification Reward (R_{sem}) addresses the semantic sufficiency of the visual perception c. To be specific, we re-prompt the same policy π_θ with only the generated perception c as a text-only proxy for the image. If the model can still derive the correct answer a^* from (c, q), the perception is considered semantically complete. Formally:

$$\hat{a} = f_\theta(c, q), \quad R_{\text{sem}}(Q, c) = \mathbb{I}[\hat{a} = a^*] \tag{2}$$

This self-reward mechanism compels the model to generate faithful and comprehensive visual descriptions, alleviating hallucinations caused by omission of information.

Geometric Verification Reward (R_{geo}) complements the semantic reward by verifying the **geometric precision** of the evidence. While R_{sem} ensures the *what* is correct, R_{geo} ensures the *where* is accurate. Let $\{\hat{b}_i\}_{i=1}^N$ be the predicted bounding box set from c, and $\{b_k^*\}_{k=1}^M$ be the ground-truth boxes. We define R_{geo} with a dual Intersection-over-Union (IoU) objective that balances recall and precision:

$$\begin{aligned} R_{\text{geo}} = & \frac{1}{2M} \sum_{k=1}^{M} \max_i \text{IoU}(b_k^*, \hat{b}_i) \\ & + \frac{1}{2N} \sum_{i=1}^{N} \max_k \text{IoU}(b_k^*, \hat{b}_i) \end{aligned} \tag{3}$$

which provides a direct, traceable supervisory signal that anchors the model's textual claims to precise spatial locations in the image, alleviating hallucinations caused by false information.

However, addressing tasks that are susceptible to hallucination and require implicit reasoning presents significant challenges that exceed the capability of a simple decompositional framework. The inherent complexity of these tasks often necessitates long-horizon reasoning, a form of "slow thinking", to arrive at a valid conclusion. A monolithic forward-pass CoT generation is often insufficient, as the model is prone to generating intermediate steps that are plausible yet fallacious, overlooking nuanced logical dependencies, and failing to detect internal inconsistencies within its own generated rationale. To address such complex reasoning scenarios, we eschew simple, linear thought sequences in favor of a more structured search framework. We extend our methodology to support multi-step, deliberative reasoning with backtracking capabilities. This enables the model to explore a diverse portfolio of reasoning paths, self-correct upon detecting errors, which are validated by a multifaceted reward signal, and iteratively refine its understanding until a verifiable solution is attained.

2.2 Search-Based Chain-of-Thought with Backtracking

While the multi-faceted reward function provides strong guidance for the reasoning process, more complex hallucination detection and safety-related tasks demand more than a single linear trace; they instead require iterative refinement. Consequently,

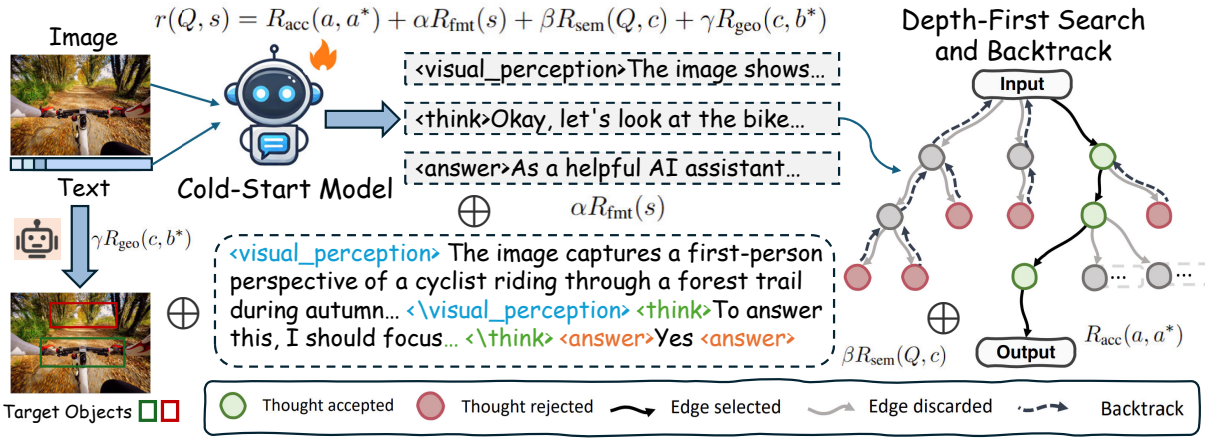

Figure 2: An overview of our VAR framework

we further reformulate the CoT generation stage as a structured search process over the reasoning trajectory space, implemented via a Depth-First Search (DFS) strategy. Within this framework, the model generates candidate reasoning steps using syntactic control tokens (<node>, <done>, <backtrack>) to extend its current path. At each step, the trajectory's validity is assessed by repurposing our reward components as validators; a failure to meet predefined thresholds for semantic consistency (R_{sem}), geometric grounding (R_{geo}), or logical coherence triggers a strategic backtrack to a previously validated ancestor node. This mechanism allows the model to systematically abandon fallacious paths and explore alternative reasoning branches until a satisfactory solution is reached or the computational budget is exhausted. This search-based formulation seamlessly integrates with our reward framework: R_{sem} and R_{geo} serve as intermediate heuristics guiding the search, R_{acc} provides the definitive signal for termination, and the format reward R_{fmt} is extended to enforce the syntactic integrity of the search protocol itself.

We formalize this search process as the construction of a Reasoning Trajectory Space, which takes the topological form of a rooted tree $G = (V, E)$. This tree-based search is specifically used to generate the reasoning trace t that forms the content of the <think>...</think> block in our final output trajectory. Each vertex $v \in V$ in this space corresponds to a semantic unit $\tau_v \in \Sigma^*$, representing a coherent proposition in a reasoning chain. A bijective label map, $\lambda : V \to L$, assigns each vertex a unique identifier from an ordered label set L (e.g., $\mathbb{N}_0$), with the root v_0 (representing the problem description) assigned the label $\lambda(v_0) = 0$.

We constitute any path from the root to a terminal vertex as a candidate CoT. The set of terminal vertices, $V_T \subset V$, is partitioned into two disjoint subsets: a set of solution vertices, V_{done}, and a set of backtracking vertices, V_{bt}. A path concluding at a vertex in V_{done} represents a complete, proposed solution, at which point the generative process for the tree terminates. Conversely, a path ending at a vertex $v_b \in V_{\text{bt}}$ triggers a state-reset, governed by a backtracking map $\beta : V_{\text{bt}} \to V$. A crucial structural constraint is imposed on this map: its codomain is restricted to the set of ancestors of the backtracking vertex, i.e., $\beta(v_b) \in \text{Ancestors}(v_b)$, ensuring that the reasoning process can only revert to a valid, previously established state within its own trajectory.

The generation of paths within this space is governed by an auto-regressive policy, $\pi_\theta(t_i|t_{<i})$, which defines a conditional probability distribution over the next token t_i given the history $t_{<i}$. The parameters θ of this policy are the subject of our learning objective. To ensure that all generated sequences are valid paths in a reasoning tree, we employ constrained decoding. This is achieved by augmenting the model's vocabulary with a set of syntactic control tokens—<node>, <backtrack>, and <done>—which delimit the reasoning steps, signal the state-reset operation, and mark the successful termination of a trajectory, respectively. At each generative step, the probability distribution π_θ is dynamically masked, effectively projecting it onto the subspace of grammatically valid continuations and thereby forcing adherence to the tree's structural rules. The construction of the full tree is thus an iterative process, where the outcome of one path generation (termination or backtracking)

determines the initial state for the next.

The construction of the complete reasoning tree, G, is therefore an iterative process, where state transitions are dictated by the terminal event of each generated path. A trajectory culminating in a <done> token transitions the process to an absorbing state, finalizing the tree's construction. Conversely, a trajectory terminating with a <backtrack> token initiates a state reset to its specified ancestor node, which then serves as the root for the subsequent path generation. We defer further discussion on the conditioning context and semantic validity to the Appendix.

Our objective is to identify a set of sufficient conditions on the policy parameters θ that ensure the generative process π_θ produces "good" search trees—defined as those that terminate efficiently and with high probability of correctness. The generative procedure, as described, follows a depth-first search strategy, with a maximum path length constrained by $T_{\max}$.

To formalize this, we introduce two critical properties that a well-behaved policy must satisfy:

Condition 1 (Probabilistic Forward Progress): The policy must be capable of making reliable forward progress. To any correct but incomplete reasoning path $v_1, \ldots, v_i$ where $i < T_{\max}$, the policy π_θ must be γ-progressive with a probability no less than $1 - \epsilon$. A policy is defined as γ-progressive if it generates a correct continuation node v_{i+1} with probability no less than γ.

Condition 2 (Reliable Trajectory Recovery): The policy must be robust to its own errors. Formally, for any incorrect reasoning path c that deviates from a correct path at node i, π_θ must induce a backtracking action to a valid ancestor node with a probability of at least $1 - \epsilon$.

In Appendix A, within the proof of the lemma, we demonstrate that these two properties suffice for the generation of an effective search tree.

3 Experiment

3.1 VAR Implementation

While end-to-end reinforcement learning (RL) is a powerful paradigm for visual grounded reasoning (VGR), its direct application from a randomly initialized policy is often computationally infeasible. The primary obstacle is the immense and unstructured nature of the search space, where a policy must learn to interleave textual reasoning with the generation of precise bounding box coordinates. In such a vast space, the sparse reward signal from a correct final answer is insufficient to guide the initial stages of exploration, leading to prohibitively long and inefficient training cycles.

To implement our proposed VAR framework, we utilized Qwen-2.5-VL-7B as the base model. The training pipeline begins with a Supervised Fine-Tuning (SFT) stage on our cold-start dataset to familiarize the model with the target output grammar. Following this initial phase, the model is further trained using Group Relative Policy Optimization (GRPO). Specifically, we trained our model, designated as VAR-7B-CI (where "CI" stands for Cold Initialization), using the LLaMA-Factory toolkit on a platform of 8*A100. The training was conducted with the AdamW optimizer, a learning rate of 5e-6, and a global batch size of 256. We employed a cosine learning rate decay schedule with a warmup ratio of 0.1.

3.2 Data Preparation

Cold Start SFT Data. Our Supervised Fine-Tuning (SFT) dataset is derived from VGR-158K, which provides pseudo-chain-of-thought annotations augmented with bounding boxes necessary for visual reasoning. To construct our initial dataset, we prompted the Qwen-2.5-VL-7B model to generate responses for each query, retaining only those samples that adhered to the target format and yielded the correct answer. This process enabled the model to rapidly adapt to the desired output grammar, establishing a robust foundation for the subsequent reinforcement learning (RL) phase.

VAR-RL-32K. To curate a dataset that prioritizes complex reasoning pathways, we filtered the VGR-158K samples, preserving only instances where the reasoning trace involved multiple bounding boxes (i.e., more than one box per trajectory). Furthermore, we incorporated the SSUI dataset, which is tailored for solving implicit reasoning safety problems requiring long-chain thought processes, and enriched it with bounding boxes via an automated annotation procedure. This culminated in a final dataset of 32,800 samples, which we denote as VAR-RL-32K, motivated by the principle that tasks involving multi-box interactions place a greater demand on spatio-temporal reasoning abilities than their single-box counterparts.

3.3 Evaluated Models and Configurations

We evaluate both open-source and closed-source MLLMs. For open-source MLLMs, recently re-

Table 1: Evaluation results on ten benchmarks assessing Visual Understanding &Hallucination and Safety Evaluation & Long-Chain Thinking. Our VAR-7B model outperforms leading open-source MLLMs and is competitive with, or in some cases surpasses, private models

	Avg	Visual Understanding & Hallucination					Safety Evaluation & Long-Chain Thinking				
		HalB	MMB	MMMU	SEED	V^*B	MGD	MSSB	SafeB	VLSB	RSB
Private Models											
Gemini-2.5-Flash	73.3	72.9	82.9	63.9	83.2	83.8	49.6	67.5	97.2	66.1	65.5
GPT-4o-1120	73.6	75.2	82.3	69.5	82.5	82.2	57.8	69.2	96.5	69.4	70.8
Gemini-2.5-Pro	78.8	76.2	85.1	68.5	86.9	92.3	55.3	73.2	98.3	75.9	76.8
Claude-3.7-Sonnet	79.6	77.3	87.2	69.2	85.3	95.5	53.7	72.1	99.5	82.3	76.2
Open-source General Models											
LLaVA-OneVision-7B	52.7	46.9	67.2	51.3	65.5	75.4	24.3	58.8	72.5	12.6	51.2
VILA-1.5-7B	51.3	45.8	68.5	51.9	63.5	72.3	16.3	52.5	69.8	14.7	57.5
Qwen2.5-VL-7B	54.3	48.3	69.2	52.8	68.2	71.2	27.9	55.4	76.2	19.0	54.8
Qwen2.5-VL-32B	60.6	48.1	75.4	54.8	69.6	87.9	43.4	55.3	87.2	26.3	58.1
InternVL3-8B	58.3	50.2	75.3	51.6	67.4	76.3	42.2	54.6	83.2	23.1	58.4
GLM-4v-9B	56.3	51.1	72.1	52.2	62.2	78.8	23.4	50.9	88.6	22.5	60.5
MiniCPM-V-2.6	53.2	47.2	68.4	50.3	63.5	74.8	32.2	48.2	75.5	16.1	56.2
Open-source Visual Reasoning Models											
DeepEyes-7B	59.6	49.2	70.6	53.8	65.2	90.1	51.5	52.2	85.2	26.6	51.7
Pixel-Reasoner-7B	58.6	35.7	69.8	53.5	66.1	89.8	48.5	54.1	86.5	27.5	54.1
VAR-7B	**72.1**	55.5	**78.5**	58.8	**79.3**	90.3	**63.1**	**74.8**	**98.5**	**43.5**	**79.1**
Δ *v.s.* Qwen2.5-VL-7B	↑ 17.8	↑ 7.2	↑ 9.3	↑ 6.0	↑ 11.1	↑ 19.1	↑ 35.2	↑ 19.4	↑ 22.3	↑ 24.5	↑ 24.3
Δ *v.s.* Qwen2.5-VL-32B	↑ 11.5	↑ 7.4	↑ 3.1	↑ 4.0	↑ 9.7	↑ 2.4	↑ 19.7	↑ 19.5	↑ 11.3	↑ 17.2	↑ 21.0

leased mainstream models are taken into consideration, which include `Qwen2.5-VL` series (Bai et al., 2025), `InternVL2` series (Chen et al., 2024), `GLM-4V` (GLM et al., 2024), `LLaVA-OneVision` series (Li et al., 2024), `MiniCPM-v2.6` (Yao et al., 2024), and `VILA` series (Lin et al., 2024). For close-source commercial MLLMs, we select `GPT-4o`, `Claude-3.7-Sonnet`, and the `Gemini` series. Furthermore, since the two most recent visual grounding reasoning models, `DeepEyes` (Zheng et al., 2025) and `Pixel-Reasoner` (Su et al., 2025), both follow a "ground-then-answer" pipeline and possess the capability to "think on image," we also include them within the scope of our comparison. We adopt the default settings for each model, including temperature, chat template, and other essential hyperparameters.

3.4 Main Result

Visual Understanding & Hallucination Evaluation. We evaluated the visual understanding and hallucination generation of VAR. For hallucination assessment, we use HalB (Guan et al., 2024), a benchmark designed to evaluate a multimodal model's ability to handle linguistic and visual illusions. To validate the model's general visual understanding capabilities, we evaluated its performance on three widely-used benchmarks: MMB (Liu et al., 2024b), MMMU (Yue et al., 2024), and SEED (Li et al., 2023a). Additionally, we conducted tests on the high-resolution V^*B (Wu and Xie, 2024) to investigate the impact of image resolution.The experimental results in Table 1 show that our method, VAR, achieves a significant 7.2% improvement on the HalB benchmark over its base model, `Qwen2.5-VL-7B`. This enhancement brings its performance remarkably close to that of the much larger `Qwen2.5-VL-32B`, demonstrating VAR's superior capability in handling both linguistic and visual illusions. Similarly, on visual understanding benchmarks, our model achieves generalizable performance improvements when compared against the `Qwen2.5-VL` series of varying scales. In high-resolution benchmark tests, our model continues to excel, significantly outperforming existing open-source models. This suggests that the "think on image" visual reasoning ability is crucial for high-resolution perception.

Safety Evaluation & Long-Chain Thinking. Our experiments are conducted on various multimodal safety benchmarks. For safety assessment, we employ MSSB (Zhou et al., 2024) to evaluate contextual safety. We also utilize three comprehensive safety suites: MGD (Gu et al., 2024), which assesses five key safety dimensions; SafeB (Ying

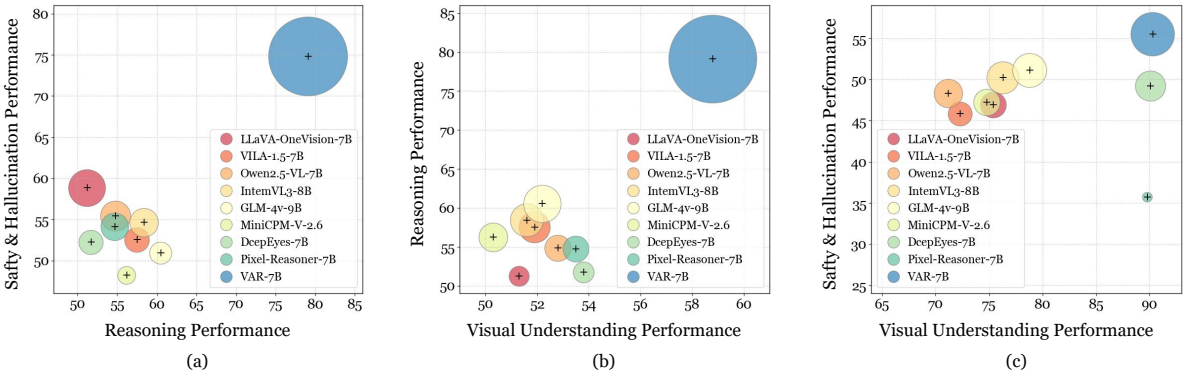

Figure 3: Correlation of Model Capabilities

et al., 2024), a comprehensive framework that evaluates MLLMs against a detailed taxonomy of 8 primary risk categories and 23 sub-categories; and VLSB (Hu et al., 2024), a reliable cross-modal benchmark structured around a safety taxonomy of 6 main categories and 19 sub-categories. To evaluate the model's long-chain reasoning capabilities, we also conduct evaluations on the RSB (Cai et al., 2025a) benchmark. The experimental results in Table 1 demonstrate the effectiveness of our proposed VAR in enhancing both the safety capabilities and the long-chain reasoning abilities of MLLMs. By applying VAR to `Qwen2.5-VL-7B`, the resulting VAR-7B achieves significant improvements across selected challenging cross-modal safety and long-chain reasoning benchmarks, with an average performance increase of 25.14%. Notably, VAR-7B exhibits superior performance even when compared to the larger `Qwen2.5-VL-32B`. Furthermore, on the MGD, MSSB, and RSB benchmarks, VAR-7B surpasses even the state of the art MLLMs.

3.5 Further Analysis

Comparative Analysis of VAR-7B's Multimodal Capabilities. As detailed in Table 2, we evaluated the comprehensive multimodal capabilities of VAR-7B by comparing it against its base model, `Qwen2.5-VL-7B`, on several conventional multimodal benchmarks. Specifically, we selected MMBench (Liu et al., 2024b), POPE, and HallusionBench (Guan et al., 2024) to assess visual-reasoning question answering (VQA) capabilities. For vision-centric question answering, we employed three benchmarks: CV-Bench, MMVP, and RealWorldQA. Document and chart understanding capabilities were evaluated using AI2D and ChartQA. We observed significant performance gains in the majority of cases, with particularly strong performance on the visual-reasoning and vision-centric benchmarks. It is noteworthy that VAR-7B outperforms the significantly larger `Qwen2.5-VL-72B` on MMBench, POPE, CV-Bench-2D, and MMVP.

Analyzing the Correlation Between Model Capabilities. In Figure 3, we conduct a systematic comparison of VAR and other open-source models, focusing on their performance in Safety & Hallucination, Visual Understanding, and Reasoning. This analysis aims to investigate the potential correlations between these capabilities. The results reveal a "decoupled" characteristic among the performance metrics on different benchmarks. For instance, while LLaVA-OneVision achieves top-tier performance in Safety & Hallucination, it lags behind peer models in the other domains. In contrast, our VAR demonstrates superior and well-rounded performance across all three areas, significantly outperforming the other models.

3.6 Ablation Study

The core contribution of VAR lies in its traceable training pipeline and the design of its backtrack mechanism. The pipeline integrates a Semantic Verification Reward (R_{sem}) and a Geometric Verification Reward (R_{geo}) into the conventional RL framework. Accordingly, we aim to evaluate the effectiveness of introducing this traceability component and the backtrack mechanism. As presented in Table 3, we conducted ablation studies on the individual components of VAR, including its cold-start initialization, the reward functions, and the backtrack mechanism.

The cold-start initialization phase is highly beneficial for visual grounding reasoning, as evidenced by the comparison between settings ① and ②. This suggests that enforcing a structured output

Table 2: VAR-7B vs. Base Model: Performance on Visual Reasoning, Vision-Centric, and Document Understanding Tasks

Capability	**Benchmark**	**Qwen2.5-VL-7B**	**VAR-7B**	**Qwen2.5-VL-72B**
Visual-Reasoning-QA	MMBench	70.3	**79.5** ↑ 9.2	78.3
	POPE	85.7	**87.5** ↑ 1.8	84.9
	HallusionBench	48.3	**55.5** ↑ 1.9	55.6
Vision-Centric-QA	CV-Bench-2D	74.0	**78.9** ↑ 4.9	77.7
	CV-Bench-3D	72.3	**79.6** ↑ 7.3	87.1
	MMVP	66.6	**75.1** ↑ 8.5	66.6
Document and chart	AI2D	**85.9**	85.7 ↓ 0.2	88.7
	ChartQA	85.5	**86.8** ↑ 1.3	89.5

Table 3: Ablations of each component of our VAR.

				Rewards			**V*B**	**HallusionBench**	**RSB**
		Cold-Start	Backtrack	$R_{\text{acc}} + R_{\text{fmt}}$	R_{sem}	R_{geo}	Acc	Acc	Acc
①	Qwen2.5-VL-7B						71.2	48.3	54.8
②	Cold-Start	✓					75.4	49.6	60.3
③	**VAR-7B**	✓	✓	✓	✓	✓	**90.3**	**55.5**	**79.1**
④	w/o Trace	✓	✓	✓			83.9	51.6	65.3
⑤	w/o Geo	✓	✓	✓	✓		86.7	53.9	72.1
⑥	w/o Sem	✓	✓	✓		✓	88.5	54.1	72.9
⑦	w/o Backtrack	✓		✓	✓	✓	87.1	52.3	68.9
⑧	Text-Only RL			✓			81.8	50.3	62.5

format for target instance bounding boxes is effective for conventional visual grounding benchmarks like V* Bench. Reasoning augmented with semantic and geometric rewards also proves effective, as demonstrated by the comparison between ③ and ④. Starting from the same cold-start checkpoint, integrating the dual rewards into the RL framework yields a significant performance boost. This improvement indicates that precise and interpretable reasoning paths are crucial for achieving optimal performance, and it highlights the value of structured reward design in complex, real-world tasks.

By comparing setting ③ with ⑤, ⑥, and ⑦, we observe that precise and complete localization is particularly important for enhancing the model's visual understanding. The performance degradation is most pronounced on V* Bench when R_{geo} is absent. In contrast, on the hallucination-focused benchmark, HallusionBench, the improvements from the semantic reward (R_{sem}) and the geometric reward (R_{geo}) are less substantial than that brought by the backtrack mechanism. This suggests that the backtrack mechanism is highly effective in mitigating hallucination. On RSBench, a benchmark requiring long-chain reasoning, the performance gains from individual components are considerably smaller than the synergistic improvement achieved when all three are combined.

The efficacy of the baseline text-only RL is less pronounced than that of visual grounding reasoning, as shown by the comparison of ③ and ⑦. While the baseline RL demonstrates value through its text-space reasoning capabilities, the performance enhancement becomes substantially more significant when visual grounding is integrated with traceable evidence. This highlights two factors: (1) contextual grounding prior to answering, which anchors the response in multimodal evidence; and (2) precise spatial localization to enhance decision-making accuracy.

4 Conlusion

We introduce Visual Attention Reasoning (VAR), a framework that reformulates linear vision-language generation as a structured search with backtracking. By enforcing traceable evidence grounding via a multi-faceted semantic and geometric reward system, VAR effectively mitigates the error propagation inherent in conventional models. Theoretical analysis and extensive experiments confirm that VAR establishes a new state-of-the-art on challenging hallucination and safety benchmarks, offering

a robust path toward trustworthy multimodal reasoning.

Limitations

VAR currently grounds reasoning via bounding boxes. While effective for object-level tasks, this representation lacks the precision required for fine-grained analysis of irregular shapes or pixel-level details. Future iterations could extend the Geometric Verification Reward (R_{geo}) to support pixel-wise segmentation masks or point-based prompts, enabling the framework to handle more nuanced visual tasks.

References

Hongjun An, Wenhan Hu, Sida Huang, Siqi Huang, Ruanjun Li, Yuanzhi Liang, Jiawei Shao, Yiliang Song, Zihan Wang, Cheng Yuan, and 1 others. 2025. Ai flow: Perspectives, scenarios, and approaches. *arXiv preprint arXiv:2506.12479*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Wei Cai, Shujuan Liu, Jian Zhao, Ziyan Shi, Yusheng Zhao, Yuchen Yuan, Tianle Zhang, Chi Zhang, and Xuelong Li. 2025a. When safe unimodal inputs collide: Optimizing reasoning chains for cross-modal safety in multimodal large language models. *arXiv preprint arXiv:2509.12060*.

Wei Cai, Jian Zhao, Yuchu Jiang, Tianle Zhang, and Xuelong Li. 2025b. Safe semantics, unsafe interpretations: Tackling implicit reasoning safety in large vision-language models. *arXiv preprint arXiv:2508.08926*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Heng Wang, Qi Han, and Yanghua Xiao. 2025. Reward shaping to mitigate reward hacking in rlhf. *arXiv preprint arXiv:2502.18770*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Tianle Gu, Zeyang Zhou, Kexin Huang, Liang Dandan, Yixu Wang, Haiquan Zhao, Yuanqi Yao, Yujiu Yang, Yan Teng, Yu Qiao, and 1 others. 2024. Mllmguard: A multi-dimensional safety evaluation suite for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:7256–7295.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. 2025. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. *arXiv preprint arXiv:2504.07086*.

Xuhao Hu, Dongrui Liu, Hao Li, Xuanjing Huang, and Jing Shao. 2024. Vlsbench: Unveiling visual leakage in multimodal safety. *arXiv preprint arXiv:2411.19939*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Yuchu Jiang, Jian Zhao, Yuchen Yuan, Tianle Zhang, Yao Huang, Yanghao Zhang, Yan Wang, Yanshu Li, Xizhong Guo, Yusheng Zhao, and 1 others. 2025. Never compromise to vulnerabilities: A comprehensive survey on ai governance. *arXiv preprint arXiv:2508.08789*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Zhimin Li, Haichao Miao, Xinyuan Yan, Valerio Pascucci, Matthew Berger, and Shusen Liu. 2025. See or recall: A sanity check for the role of vision in solving visualization question answer tasks with multimodal llms. *arXiv preprint arXiv:2504.09809*.

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*, pages 26689–26699.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, and 1 others. 2025. Spurious rewards: Rethinking training signals in rlvr. *arXiv preprint arXiv:2506.10947*.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.

Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Language prior is not the only shortcut: A benchmark for shortcut learning in vqa. *arXiv preprint arXiv:2210.04692*.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. 2024. Chain of thoughtlessness? an analysis of cot in planning. *Advances in Neural Information Processing Systems*, 37:29106–29141.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. 2025. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Penghao Wu and Saining Xie. 2024. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. 2025. Are reasoning models more prone to hallucination? *arXiv preprint arXiv:2505.23646*.

Zonghao Ying, Aishan Liu, Siyuan Liang, Lei Huang, Jinyang Guo, Wenbo Zhou, Xianglong Liu, and Dacheng Tao. 2024. Safebench: A safety evaluation framework for multimodal large language models. *arXiv preprint arXiv:2410.18927*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*.

Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Anderson Compalas, Dawn Song, and Xin Eric Wang. 2024. Multimodal situational safety. *arXiv preprint arXiv:2410.06172*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Appendices

A Lemma Proof

Lemma 1 (Sufficient Conditions for High-Probability Trajectory Generation)**.** *Let $\gamma \in (0, 1)$ and $\delta \in (0, 1/2)$ be constants. Define error tolerance ϵ and exploration budget B as:*

$$\epsilon = \frac{\gamma\delta}{2T_{\max}} \quad \textit{and} \quad B = \left\lceil \frac{\ln(T_{\max}/\delta)}{\gamma} \right\rceil. \quad (4)$$

*Suppose a policy π_θ satisfies the **Probabilistic Forward Progress** and **Reliable Trajectory Recovery** conditions with these parameters, and we restrict each node expansion to at most B generative attempts. Then, the probability that π_θ generates a correct, complete Chain-of-Thought (CoT) is at least $1 - 2\delta$. Furthermore, the total number of backtracking leaves in the search tree is bounded by $(B - 1)(T_{\max} - 1)$.*

Proof. Let P_{adv} be the probability of successfully advancing from a correct node v_i to the next correct node v_{i+1} within B attempts. We lower-bound this probability by summing over the mutually exclusive events of succeeding at the k-th attempt ($1 \le k \le B$).

Success at the k-th attempt requires $k-1$ consecutive "successful failures" followed by one "direct success."

- A "direct success" occurs with probability $p_s \ge \gamma(1-\epsilon)$, accounting for the policy being γ-progressive (with probability at least $1 - \epsilon$) and subsequently generating the correct token (with probability at least γ).
- A "successful failure" occurs when an incorrect token is generated but the trajectory is correctly recovered via backtracking. This occurs with probability $p_f = (1 - \gamma)(1 - \epsilon)$, as recovery is guaranteed with probability $1 - \epsilon$.

The probability of advancing, P_{adv}, is the sum of a geometric series:

$$P_{\text{adv}} = \sum_{k=1}^{B} (p_f)^{k-1} p_s = p_s \frac{1 - p_f^B}{1 - p_f}. \quad (5)$$

We first bound the denominator $1 - p_f = 1 - (1 - \gamma)(1 - \epsilon) = \gamma + \epsilon - \gamma\epsilon$. Since $p_s \ge \gamma(1 - \epsilon)$, the ratio $\frac{p_s}{1-p_f}$ is $\frac{\gamma(1-\epsilon)}{\gamma+\epsilon-\gamma\epsilon}$. Let $X = \gamma(1 - \epsilon)$. Then this ratio is $\frac{X}{X+\epsilon} \ge 1 - \frac{\epsilon}{X} = 1 - \frac{\epsilon}{\gamma(1-\epsilon)} \ge 1 - \frac{2\epsilon}{\gamma}$ for sufficiently small ϵ.

Next, we bound the term $(1 - p_f^B)$. Using the inequality $1-x \le e^{-x}$, we have $p_f^B = ((1-\gamma)(1-\epsilon))^B \le (e^{-\gamma}e^{-\epsilon})^B = e^{-(\gamma+\epsilon)B}$. Consequently, $P_{\text{adv}} \ge (1 - \frac{2\epsilon}{\gamma})(1 - e^{-(\gamma+\epsilon)B})$.

By our definition of B, $\gamma B \ge \ln(T_{\max}/\delta)$, which implies $e^{-\gamma B} \le \delta/T_{\max}$. Since $e^{-\epsilon B} \le 1$, it follows that $e^{-(\gamma+\epsilon)B} \le \delta/T_{\max}$. Thus, $P_{\text{adv}} \ge (1 - \frac{2\epsilon}{\gamma})(1 - \frac{\delta}{T_{\max}})$.

Let T be the length of a correct reasoning chain ($T < T_{\max}$). The probability of successfully generating the entire chain, $P_{\text{succ}}(T)$, is $(P_{\text{adv}})^T$.

$$P_{\text{succ}}(T) \ge \left[\left(1 - \frac{2\epsilon}{\gamma}\right) \left(1 - \frac{\delta}{T_{\max}}\right) \right]^T. \quad (6)$$

Using the inequality $(1 - x)(1 - y) \ge 1 - x - y$ for small positive x, y:

$$P_{\text{succ}}(T) \ge \left(1 - \frac{2\epsilon}{\gamma} - \frac{\delta}{T_{\max}}\right)^T. \quad (7)$$

Using the inequality $(1 - x)^n \ge 1 - nx$:

$$\begin{aligned} P_{\text{succ}}(T) &\ge 1 - T\left(\frac{2\epsilon}{\gamma} + \frac{\delta}{T_{\max}}\right) \\ &= 1 - \left(\frac{2T\epsilon}{\gamma} + \frac{T\delta}{T_{\max}}\right). \end{aligned} \quad (8)$$

Substituting $\epsilon = \frac{\gamma\delta}{2T_{\max}}$:

$$\begin{aligned} P_{\text{succ}}(T) &\ge 1 - \left(\frac{2T(\gamma\delta/2T_{\max})}{\gamma} + \frac{T\delta}{T_{\max}}\right) \\ &= 1 - \left(\frac{T\delta}{T_{\max}} + \frac{T\delta}{T_{\max}}\right) \\ &= 1 - \frac{2T\delta}{T_{\max}}. \end{aligned} \quad (9)$$

Since $T < T_{\max}$, the ratio $\frac{T}{T_{\max}} < 1$. Thus, $P_{\text{succ}}(T) > 1 - 2\delta$. The number of backtracking leaves is at most $B - 1$ for each of the $T_{\max} - 1$ potential intermediate steps, yielding the stated complexity bound. □

Remark 1 (Relaxing the Backtracking Precision)**.** *Lemma 1 assumes the backtrack operation lands on the exact optimal node $\beta(c)$. We can relax this assumption. For an incorrect sequence c, we can tolerate **undershooting**, where the policy backtracks to a node i slightly prior to the optimal one ($i > \beta(c)$). This is a benign error, requiring only a constant number of additional steps to re-derive known-correct work, and does not compromise the overall convergence guarantee. In contrast, **overshooting** ($i < \beta(c)$) must remain a low-probability*

event ($< \epsilon$). An overshoot is a costly error that discards significant progress, analogous to a catastrophic state reset (reminiscent of sliding down a chute in "Chutes and Ladders"). This asymmetry is advantageous: we can design our learning objective to be highly intolerant of overshoots, even if it implies accepting a higher rate of benign undershoots.

B The Conditioning Context and Semantic Validity

B.1 On the Conditioning Context

When generating a path from a given vertex v, the choice of conditioning context for the policy π_θ represents a critical design decision. One may condition on the entire history of the generated tree G_t. While providing maximal information regarding prior failed explorations, this introduces non-stationarity that can lead to distributional drift during learning. In this work, we adopt a constrained Markovian assumption: the generation of a new path depends exclusively on the linear sequence of vertices from the root v_0 to the current starting vertex v. From an implementation standpoint, this choice enhances stability and is computationally efficient, as it allows for optimization via standard key-value (KV) caching mechanisms.

B.2 On Semantic Validity

Our constrained decoding mechanism ensures only the *syntactic correctness* of the generated tree structure. To ensure *semantic validity*—i.e., that each reasoning step is logically and factually sound—we assume the availability of a Validation Oracle. As argued in prior work, the problem of validating a given reasoning step belongs to a significantly lower complexity class than generating it *de novo*. This oracle is therefore responsible for verifying that each proposition τ_v logically follows from its parent, $\tau_{\text{parent}(v)}$, along any path leading to a V_{done} vertex.

C The Shortest Reasoning Path Example

Lemma 2 (Characterization of the Optimal Policy via Parity)**.** *For any permutation $\pi \in S_n$ and input $x \in \{\pm 1\}^n$, let $p^*(x, \pi)$ denote the optimal (shortest) path from source s to sink t. Let $f_\pi(x)$ be defined by the first vertex on this path after s, such that $f_\pi(x) = +1$ if the vertex is a_0 and $f_\pi(x) = -1$ if it is b_0. Then, this function is given by:*

$$f_\pi(x) = \prod_{1 \le j \le n,\, j \text{ is odd}} x_{\pi(j)}. \tag{10}$$

Proof. The total cost of any path p from s to t is determined solely by the cumulative costs incurred at the odd-indexed layers $j \in \{1, 3, \ldots, n-1\}$, as all edges at even-indexed layers have zero weight.

At each odd layer j, an agent must decide whether to maintain its current row (transitioning from a_{j-1} to a_j) or switch rows (transitioning from a_{j-1} to b_j). The cost of this decision is conditional on the input bit $x_{\pi(j)}$. The locally optimal action at layer j is always the unique action incurring zero cost. This is achieved by switching rows if and only if $x_{\pi(j)} = -1$.

Consequently, the total number of row switches along any optimal, zero-cost path is precisely equal to the count of -1's in the odd-indexed positions of x under permutation π. Let this count be $T(x, \pi) = |\{j \le n : j \text{ is odd and } x_{\pi(j)} = -1\}|$.

For a path to be optimal, it must have a total cost of zero, requiring it to terminate at vertex a_n before the final transition to the sink t. A path starting at a_0 can only reach a_n after an even number of row switches. Conversely, a path starting at b_0 can only reach a_n after an odd number of row switches.

Therefore, for the total path cost to be zero, the optimal starting vertex must be a_0 if the required number of switches, $T(x, \pi)$, is even, and b_0 if $T(x, \pi)$ is odd. This condition on the optimal initial action is equivalent to stating $f_\pi(x) = (-1)^{T(x,\pi)}$.

The final identity, $(-1)^{T(x,\pi)} = \prod_{j \text{ is odd}} x_{\pi(j)}$, follows directly from the definition of $T(x, \pi)$, as each -1 at an odd position contributes a factor of -1 to the product, while each $+1$ contributes a factor of 1. This concludes the proof. □