FOCOT: A Unified Framework of FOT for Multimodal VOCOT

Anonymous ACL submission

Abstract

Multimodal Chain-of-Thought (CoT) improves interpretability and spatial grounding by fusing vision and language. However, it still reasons along a single trajectory, ignoring complementary information from different perspectives. Inspired by the success of Forest-of-Thought (FOT) in long-text reasoning, we reframe object-aware CoT as a multibranch, tree-structured exploration that allows the model to traverse diverse reasoning paths under complex vision-language scenarios and to converge toward an optimal thinking trajectory. Specifically, we expand textual and visual cues into distinct reasoning branches that capture complementary perspectives, and introduce node-level collaboration to enable cross-branch evidence exchange and soft aggregation of partial signals. To further curb hallucinations, we introduce node-level contrastive decoding, extending discriminative decoding to operate on the entire reasoning forest structure. Extensive experiments demonstrate that our framework, Forest-of-Chain-of-Thought (FOCOT), achieves superior performance across multiple benchmarks, significantly surpassing VOCOT, standard FOT, and other strong CoT baselines. Moreover, ablation studies further confirm the effectiveness of each component within our framework. The code is available in the Appendix.

1 Introduction

Recent advances in Multimodal Chain-of-Thought (MMCoT) have significantly improved reasoning interpretability by fusing visual and linguistic representations (Shao et al., 2024; Wang et al., 2023b). However, traditional MMCoT models predominantly rely on a *linear decoding paradigm*, where each reasoning step depends deterministically on its predecessor (Wei et al., 2022). This sequential nature renders the process vulnerable to **cascading failures**: an early visual misinterpretation propagates through the trajectory, leading to semantically coherent but factually incorrect hallucinations (Wang et al., 2023c; Yao et al., 2023a). Furthermore, single-path search inherently restricts exploration, ignoring alternative reasoning routes that could potentially rectify early perceptual errors.

To address these limitations, we propose **Forest-of-Chain-of-Thought (FOCOT)**, extending the Forest-of-Thought (FOT) paradigm (Bi et al., 2024) to the multimodal domain. In FOCOT, reasoning unfolds as a tree-structured exploration where nodes correspond to object-grounded steps anchored in specific image regions. By pursuing multiple reasoning trajectories in parallel, the model can explore diverse interpretations of the visual scene and converge toward an optimal path.

A critical challenge in long-form multimodal reasoning is the tendency to drift toward linguistic priors while neglecting visual evidence. While contrastive decoding (Li et al., 2022; Chuang et al., 2023) has addressed text degeneration at the *token level*, it is insufficient for pruning flawed *reasoning branches* in complex vision-language tasks. Consequently, we introduce **Node-level Contrastive Decoding (NCD)**, which operates at the semantic level by dynamically contrasting "strong" and "weak" reasoning contexts. NCD re-ranks candidate nodes to penalize visually-unsupported paths, thereby suppressing hallucinations across entire semantically-competing nodes (Fig. 1).

Extensive experiments on **GQA**, **POPE**, and **AMBER** benchmarks demonstrate that FOCOT consistently outperforms strong baselines like CoT and VOCOT across various backbones (Table 1). Our main contributions are:

- **FOCOT Framework:** A tree-structured multimodal reasoning paradigm that enables

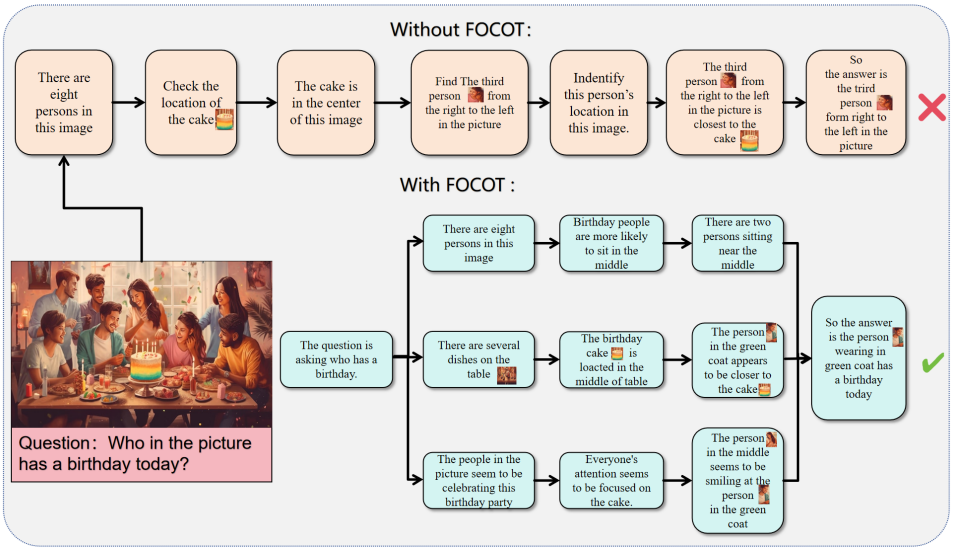


Figure 1: Example of our object-grounded forest-of-thought (FOT).

parallel exploration to overcome the brittleness of single-trajectory decoding.

- **Node-level Contrastive Decoding:** A novel semantic-level contrastive mechanism that suppresses hallucinations by re-ranking reasoning branches based on visual grounding.

- **Empirical Validation:** Extensive results confirming that our grounded collaboration and NCD are complementary, achieving state-of-the-art performance on reasoning and hallucination benchmarks.

2 Related Work

2.1 Chain-of-Thought (CoT) and Forest-of-Thought (FOT) Reasoning

Chain-of-Thought (CoT) prompting elicits intermediate steps to improve reasoning transparency (Wei et al., 2022; Kojima et al., 2022). To address the brittleness of a single trace, *self-consistency* introduced sampling multiple chains with a majority vote (Wang et al., 2023c), while *least-to-most prompting* focused on structured problem decomposition (Zhou et al., 2023).

A parallel line of work enhances reasoning by integrating external computation. *Program-Aided Language* (PAL) and *Program-of-Thoughts* (PoT) externalize symbolic or numeric steps to code execution (Gao et al., 2023; Chen et al., 2023b). Others, like *STaR*, focus on improving the rationales themselves by bootstrapping from model-generated solutions (Zelikman et al., 2022).

The paradigm then evolved beyond linear traces to non-linear exploration. *Tree-of-Thought* (ToT) casts inference as a search over branching "thoughts" (Yao et al., 2023a), later generalized to graph structures (GoT) (Besta et al., 2024). Agent-style frameworks such as *ReAct* and *Reflexion* further interleaved reasoning with tool use and iterative self-feedback (Yao et al., 2023c; Shinn et al., 2023).

Forest-of-Thought (FOT) extends this trajectory by scaling test-time compute via *parallel* exploration over multiple reasoning trees, broadening evidence coverage instead of early commitment to a single chain (Bi et al., 2024). While this parallel search improves robustness, the branches in standard FOT operate independently and lack mechanisms for explicit visual grounding or reusing partially correct informationlimitations we address in this work.

2.2 Visual Grounding and VOCOT

For multimodal tasks such as VQA, interpretability requires that intermediate steps be *spatially* aligned with the image, not merely logically coherent. VOCOT advances this goal by introducing an object-centric, coordinate-level chain-of-thought format and instruction data that anchor each step to concrete regions (Wang et al., 2023b). Visual-CoT expands the space of grounded rationales with a comprehensive dataset and benchmark for multimodal CoT, encouraging step-by-step reasoning tied to visual content (Shao et al., 2024). Argus further explores grounded chains for vision-

centric reasoning, emphasizing explicit region references and compositional cues (Man et al., 2025). Building on these insights, our approach maintains coordinate-level grounding throughout the search, ensuring that every branch and node remains anchored to the image as multi-branch exploration proceeds in parallel.

2.3 Contrastive Decoding for Reliable Generation

Contrastive Decoding selects outputs that are preferred by a strong model but *not* by a weaker one, improving coherence and reducing degeneracy at inference time without additional training (Li et al., 2022). Contrastive Search (SimCTG) balances relevance and diversity by contrasting token scores against a cache of contextual representations, mitigating repetition and blandness (Su et al., 2022). DoLa contrasts logits from different transformer layers to surface more factual continuations and suppress hallucinations (Chuang et al., 2023). Inspired by these ideas, we incorporate contrastive principles into our branch selection: candidates that are better supported by visual evidenceand less favored by contrastive baselinesare prioritized, which sharpens discrimination among branches and reduces unsupported claims.

3 Methodology

In this section, we present our proposed framework, which adapts Forest-of-Thought (FOT) (Bi et al., 2024) for complex multimodal reasoning. To achieve this, our framework introduces two novel mechanisms: *grounded node-level collaboration* and *node-level contrastive decoding*. The design of these mechanisms builds upon principles of visually-grounded reasoning (Wang et al., 2023b) and degeneration mitigation strategies, such as contrastive decoding (Li et al., 2022). The overall architecture is illustrated in Figure 2.

3.1 Overview of the Framework

Figure 2 illustrates the **FOCOT** architecture. Given an image I and a question Q, FOCOT formulates multimodal reasoning as a *tree-structured search process* consisting of four stages:

First, inputs are processed via **Vision and Text Encoders** and fused through **Cross-Attention** to form grounded representations. Second, the model initiates multiple reasoning branches via **Monte Carlo Node Selection**, instantiating diverse trajectories that capture different scene interpretations. Each branch evolves through step-wise inference, anchored by object-specific bounding boxes to ensure spatial grounding.

Third, to enhance robustness, the **Grounded Node-level Collaboration** module enables information exchange between branches, allowing non-selected nodes to provide auxiliary evidence. Finally, **Node-level Contrastive Decoding (NCD)** re-ranks candidates by contrasting "strong" and "weak" node contexts (Sec. 3.6), effectively suppressing hallucinations. The final answer is synthesized by aggregating evidence across the refined forest, balancing exploratory breadth with perceptual precision.

3.2 Encoding

Vision Encoding. We utilize CLIP ViT-L/14 $E_v(\cdot)$ to extract region-level visual features from image $\mathbf{I}$, providing the grounding foundation:

$$\mathbf{V} = E_v(\mathbf{I}) = \mathbf{v}ii = 1^{N_v}, \quad \mathbf{v}i \in \mathbb{R}^{d_v}. \quad (1)$$

Textual Encoding. Following VOCOT, we employ a LLaMA-2-7B backbone $E_t(\cdot)$. To enhance reasoning diversity, the question $\mathbf{Q}$ is augmented with multi-aspect prompts (e.g., causal, comparative) before encoding:

$$\mathbf{T} = E_t(\mathbf{Q}aug) = \mathbf{t}jj = 1^{N_q}, \quad \mathbf{t}_j \in \mathbb{R}^{d_t}. \quad (2)$$

These embeddings $\mathbf{V}$ and $\mathbf{T}$ serve as the initial representations for cross-modal alignment.

3.3 Cross-Modal Fusion with Grounding Tokens

To enable fine-grained visuallinguistic alignment, FOCOT employs a multi-head cross-attention mechanism:

$$\tilde{\mathbf{T}} = \text{CrossAttn}(\mathbf{T}, \mathbf{V}, \mathbf{V}) = \tilde{\mathbf{t}}jj = 1^{N_q}. \quad (3)$$

Grounding Tokens. To explicitly bind reasoning steps to image regions, we incorporate coordinate tokens. For each region r_i, a token $\langle\text{coor}\rangle(x_1^i, y_1^i, x_2^i, y_2^i)\langle/\text{coor}\rangle$ is appended to the language stream, where (x, y) represent normalized bounding box coordinates. These act as spatial anchors, ensuring reasoning remains grounded at the coordinate level.

Multimodal Context. The grounded representations are projected into a unified space $\mathbf{H}$:

$$\mathbf{H} = \text{Proj}([\tilde{\mathbf{T}}; \mathbf{V}]). \quad (4)$$

This context $\mathbf{H}$ integrates linguistic cues with spatial evidence, serving as the root for subsequent tree-structured inference (§3.4).

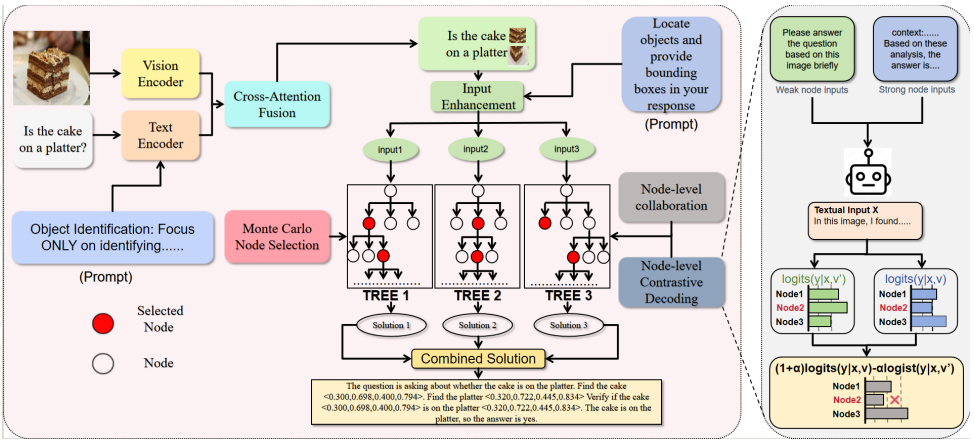


Figure 2: Overview of the proposed multimodal FOT pipeline.

3.4 Multimodal Forest-of-Thought (FOCOT)

Recap of FOT. Forest-of-Thought (FOT) (Bi et al., 2024) expands a *forest* of reasoning states rather than a single chain. Starting from h_0, the model iteratively: (i) proposes K candidate nodes via **Monte Carlo Node Selection** (stochastic sampling with look-ahead rollouts); (ii) scores nodes via a utility function s_i balancing local likelihood and global progress; and (iii) prunes the frontier to maintain efficiency.

Multimodal Grounding and Initialization. FOCOT extends FOT by representing each node h as a multimodal state. We augment textual rationales with **coordinate tokens** (§3.3), enabling nodes to *cite* image regions. To ensure initial forest diversity, we instantiate branches using targeted prompts focusing on: (1) *Visual Element Analysis* (perception), (2) *Functional Inference* (relations), and (3) *Hypothesis-Driven Deduction* (verification).

Structural Robustness. Unlike standard MM-CoT, FOCOT's gains stem from its structural mechanisms rather than prompt engineering. Diversity is elicited by prompts, but the reasoning space is navigated and refined by **Grounded Node-level Collaboration** (Alg. 1) and **Node-level Contrastive Decoding** (§3.6). The utility function s_i is specifically adapted to reward valid spatial citations, ensuring each partial thought remains visually anchored.

3.5 Grounded Node-Level Collaboration

Unlike conventional FOT (Bi et al., 2024) which treats reasoning branches independently, we introduce **Grounded Node-Level Collaboration** to reframe the forest as a collaborative graph $G = (V, E)$. This enables cross-branch information exchange to enhance global consistency (see Alg. 1).

Each node $i \in V$ represents a state $\text{Node}_i = (r_i, p_i, E_i)$ containing reasoning text r_i, a prediction p_i, and visual evidence E_i. We quantify node relationships via a weighted edge w_{ij}, defined by Semantic Similarity (S_{ij}) and Evidence Overlap (O_{ij}):

$$w_{ij} = \left(\frac{h(r_i) \cdot h(r_j)}{\|h(r_i)\| \|h(r_j)\|} \right)^{\lambda_S} \cdot \left(\frac{|T(r_i) \cap T(r_j)|}{|T(r_i) \cup T(r_j)|} \right)^{\lambda_O} \tag{5}$$

where $h(\cdot)$ are reasoning embeddings and $T(\cdot)$ are evidence terms. An edge is instantiated only if $w_{ij} \geq \tau_{\text{conn}}$, creating a sparse graph. The base confidence c_i is then refined into a **collaborative score** s'_i:

$$s'_i = c_i + \beta \sum_{j \in N(i)} \hat{w}_{ij} c_j - \gamma P_i \tag{6}$$

where $N(i)$ is the neighborhood of i, $\hat{w}_{ij}$ is the normalized weight, and P_i penalizes contradictions between p_i and its neighbors. This refined score s'_i guides the subsequent FOT search, prioritizing paths that are both internally consistent and externally supported by neighboring evidence.

3.6 Node-Level Contrastive Decoding

To mitigate visual hallucinations and linguistic drift, we introduce **Node-level Contrastive Decoding (NCD)**. Unlike standard contrastive decoding (Li et al., 2022) that operates on tokens, NCD refines node selection by contrasting reasoning against strong and weak semantic contexts.

First, we define a **strong context** $\mathcal{C}_s$ from top-k high-confidence nodes and a **weak context** $\mathcal{C}_w$

Algorithm 1: Grounded Node-Level Collaboration

Input: Nodes $\{\nu_i\}$ with (r_i, p_i, E_i); embeddings $h(r_i)$; evidence terms $T(r_i)$; hyperparams $\lambda_S, \lambda_O, \tau_{\text{conn}}, \beta, \gamma$
Output: Collaborative scores $\{s_i^{\text{coll}}\}$ and normalized edge weights $\{\hat{w}_{ij}\}$

Base confidences
foreach ν_i **do**
  $c_i \leftarrow S_{\text{len}}(r_i) + B_{\text{pred}}(p_i)$
end

Build sparse collaboration graph
foreach *pair* (i, j), $i \neq j$ **do**
  $S_{ij} \leftarrow \dfrac{h(r_i) \cdot h(r_j)}{\|h(r_i)\| \, \|h(r_j)\|}$
  $O_{ij} \leftarrow \dfrac{|T(r_i) \cap T(r_j)|}{|T(r_i) \cup T(r_j)|}$
  $w_{ij} \leftarrow (S_{ij})^{\lambda_S} \cdot (O_{ij})^{\lambda_O}$
  if $w_{ij} \geq \tau_{\text{conn}}$ **then**
    add undirected edge $i \leftrightarrow j$ with weight w_{ij}
  end
end
foreach i *with neighbors* $N(i)$ **do**
  $\hat{w}_{ij} \leftarrow \dfrac{w_{ij}}{\sum_{k \in N(i)} w_{ik}} \quad \forall j \in N(i)$
end

Conflict penalty and collaborative score
foreach i **do**
  $P_i \leftarrow \sum_{j \in N(i)} \hat{w}_{ij} \, \mathbf{1}[\text{Contradict}(p_i, p_j, E_i, E_j)]$
  $s_i^{\text{coll}} \leftarrow c_i + \beta \sum_{j \in N(i)} \hat{w}_{ij} \, c_j - \gamma \, P_i$
end
return $\{s_i^{\text{coll}}\}, \{\hat{w}_{ij}\}$

from bottom-m nodes. Node confidence is computed as $C(\nu_i) = S_{\text{len}}(r_i) + B_{\text{pred}}(p_i)$, rewarding informative reasoning and valid predictions. We then compute a contrastive score Δ_{CD} for reasoning r_i to ensure it is probable under $\mathcal{C}_s$ but improbable under $\mathcal{C}_w$:

$$\Delta_{\text{CD}}(r_i) = \sum_{t=1}^{|r_i|} [(1+\alpha) \log p(t|\mathcal{C}_s, r_{<t}) - \alpha \log p(t|\mathcal{C}_w, r_{<t})] \tag{7}$$

where α is the penalty hyperparameter. Finally, the base utility s_i is adjusted to produce a refined score $\hat{s}_i = s_i + \beta \Delta_{\text{CD}}(r_i)$, where β balances the contrastive evaluation. This training-free mechanism favors nodes supported by high-quality peers while suppressing uncertain or hallucinated paths.

4 Experiment

4.1 Experimental Setup

We evaluate our framework on three key datasets: **GQA** (Hudson and Manning, 2019) for compositional reasoning, **POPE** (Lin et al., 2014) for robustness against object hallucinations, and **AMBER** (Wang et al., 2023a) for open-ended generative reasoning.

Backbone Models. Our primary implementation and ablation studies are based on **Volcano-7B** (Lee et al., 2024). To demonstrate model-agnostic effectiveness, we extend evaluations to **MiniGPT-v2-7B** (Chen et al., 2023a) and **LLaVA-1.6-7B** (Liu et al., 2024), reporting GQA results to verify generalizability. Furthermore, to examine the **scalability** of FOCOT on advanced reasoning backbones, we evaluate our framework on the **DeepSeek-R1-Distill-Qwen-14B** model (DeepSeek-AI et al., 2025), which leverages large-scale reinforcement learning and distillation to enhance logical inference.

Implementation Details. Unless specified, we employ $K = 3$ reasoning branches with a search depth of $S = 3$. Hyperparameters are set to $\tau = 0.7$, $\lambda = 0.5$, and $\beta = 0.3$ based on empirical tuning. All experiments were conducted on NVIDIA H20 GPUs. Detailed evaluation drivers, prompt templates, and hardware configurations are provided in the **Supplementary Material**.

4.2 Evaluation Methodology

To overcome the limitations of exact string matching, we employ **Qwen-turbo** as a large-scale evaluator model (Bai et al., 2023), following recent practices in model-based evaluation. For datasets with ground truth (GQA, POPE for accuracy), both the baseline output and our FOT-enhanced output are compared against the gold answer. For open-ended tasks (Amber), the evaluator scores responses along human-interpretable dimensions.

Evaluation Dimensions. For GQA, POPE, and AMBER, we adopt four complementary, human-interpretable dimensions for each benchmark. For POPE, we additionally report its standard exact-match classification accuracy, as its final answers are strictly binary (`yes`/`no`).

All dimensions are scored on a 0–1 scale (with two decimal places) and summed into a total score (maximum 4.00). The batch evaluation pipeline collects per-dimension scores into JSON for analysis. **A detailed breakdown of all 12 dimensions, along with the evaluation driver algorithm, is available in the Appendix.**

4.3 Baselines

We compare our method against a wide range of baselines, which are grouped into several cate-

Table 1: Main experimental results. **(Top)** Results across GQA, AMBER, and POPE on the Volcano-7B backbone. FOCOT consistently outperforms all baselines. **(Bottom)** Generalizability results on GQA for MiniGPT-v2-7B and LLaVA-1.6-7B backbones.

	GQA (0–1)					AMBER (0–1)					POPE Evaluator (0–1)					POPE
Model	**Obj.**	**Log.**	**Ima.**	**Thi.**	**Total**	**Rea.**	**Vis.**	**Kno.**	**Res.**	**Total**	**Acc.**	**Log.**	**Ima.**	**Thi.**	**Total**	**Acc.**
Random Guess	0.18	0.22	0.19	0.20	0.79	0.22	0.31	0.20	0.29	1.02	0.29	0.31	0.28	0.30	1.18	24.68
Majority Answer	0.28	0.31	0.29	0.33	1.21	0.33	0.39	0.31	0.37	1.40	0.38	0.39	0.37	0.40	1.54	40.35
Vanilla LLM	0.39	0.41	0.43	0.47	1.70	0.42	0.45	0.35	0.46	1.68	0.47	0.48	0.46	0.49	1.90	43.12
Standard CoT	0.52	0.49	0.51	0.65	2.17	0.52	0.53	0.41	0.60	2.06	0.57	0.58	0.52	0.66	2.33	44.20
Self-Cons. CoT	0.61	0.63	0.58	0.70	2.52	0.59	0.61	0.45	0.63	2.28	0.61	0.67	0.59	0.70	2.57	45.86
ToT	0.65	0.74	0.63	0.76	2.78	0.65	0.67	0.48	0.68	2.48	0.66	0.74	0.63	0.77	2.80	46.75
ReAct	0.63	0.70	0.60	0.72	2.65	0.63	0.65	0.47	0.66	2.41	0.65	0.70	0.61	0.75	2.71	46.20
RAD	0.64	0.71	0.61	0.74	2.70	0.64	0.66	0.48	0.67	2.45	0.66	0.72	0.62	0.76	2.76	46.48
Reflexion	0.67	0.75	0.64	0.77	2.83	0.67	0.69	0.50	0.70	2.56	0.68	0.76	0.65	0.79	2.88	47.33
Standard FOT	0.64	0.88	0.62	0.72	2.86	0.70	0.76	**0.63**	0.74	2.80	0.71	0.89	0.72	0.80	3.12	47.92
FOT (indep. vot.)	0.63	0.82	0.64	0.77	2.86	0.69	0.74	0.57	0.73	2.73	0.70	0.84	0.71	0.79	3.04	48.20
VOCOT	0.59	0.68	0.54	0.62	2.43	0.63	**0.79**	0.55	0.77	2.74	0.63	0.72	0.74	0.75	2.84	45.43
FOT (no VOCOT)	0.71	0.79	0.66	0.74	2.90	0.72	0.75	0.59	0.75	2.81	0.72	0.83	0.76	0.80	3.11	49.32
FOCOT (Ours)	**0.83**	**0.91**	**0.74**	**0.82**	**3.30**	**0.80**	**0.79**	**0.63**	**0.78**	**3.00**	**0.83**	**0.91**	**0.88**	**0.89**	**3.51**	**54.00**

	MiniGPT-v2-7B (GQA, 0–1)					LLaVA-1.6-7B (GQA, 0–1)				
Model/Baseline	**Obj.**	**Log.**	**Ima.**	**Thi.**	**Total**	**Obj.**	**Log.**	**Ima.**	**Thi.**	**Total**
Random Guess	0.16	0.21	0.19	0.22	0.78	0.19	0.21	0.19	0.22	0.81
Majority Answer	0.27	0.30	0.30	0.33	1.20	0.31	0.31	0.28	0.32	1.22
Vanilla LLM	0.37	0.38	0.41	0.45	1.61	0.42	0.43	0.45	0.49	1.79
Standard CoT	0.50	0.46	0.49	0.63	2.08	0.55	0.51	0.53	0.67	2.26
Self-Cons. CoT	0.59	0.60	0.56	0.68	2.43	0.64	0.65	0.60	0.72	2.61
ToT	0.63	0.71	0.61	0.74	2.69	0.68	0.76	0.65	0.78	2.87
ReAct	0.61	0.67	0.58	0.70	2.56	0.66	0.72	0.62	0.74	2.74
RAD	0.62	0.68	0.59	0.72	2.61	0.67	0.73	0.63	0.76	2.79
Reflexion	0.65	0.72	0.62	0.75	2.74	0.70	0.77	0.66	0.79	2.92
Standard FOT	0.62	0.85	0.60	0.70	2.77	0.67	0.90	0.64	0.74	2.95
FOT (indep. vot.)	0.61	0.79	0.62	0.75	2.77	0.66	0.84	0.66	0.79	2.95
VOCOT	0.57	0.65	0.52	0.60	2.34	0.62	0.70	0.56	0.64	2.52
FOT (no VOCOT)	0.69	0.76	0.64	0.72	2.81	0.72	0.81	0.68	0.76	2.97
FOCOT (Ours)	**0.79**	**0.88**	**0.72**	**0.80**	**3.19**	**0.84**	**0.88**	**0.75**	**0.84**	**3.31**

Table 2: Experimental results on the **DeepSeek-R1-Distill-Qwen-14B** backbone. This table complements the main results by evaluating the scaling performance and generalizability of our proposed FOCOT framework on a larger distilled reasoning model.

	GQA (0–1)					AMBER (0–1)					POPE Evaluator (0–1)					POPE
Model	**Obj.**	**Log.**	**Ima.**	**Thi.**	**Total**	**Rea.**	**Vis.**	**Kno.**	**Res.**	**Total**	**Acc.**	**Log.**	**Ima.**	**Thi.**	**Total**	**Acc.**
Vanilla LLM	0.44	0.46	0.48	0.52	1.90	0.47	0.50	0.40	0.51	1.88	0.52	0.53	0.51	0.54	2.10	46.50
Standard CoT	0.57	0.54	0.56	0.70	2.37	0.57	0.58	0.46	0.65	2.26	0.62	0.63	0.57	0.71	2.53	47.60
Self-Cons. CoT	0.66	0.68	0.63	0.75	2.72	0.64	0.66	0.50	0.68	2.48	0.66	0.72	0.64	0.75	2.77	49.10
ToT	0.69	0.79	0.67	0.81	2.96	0.69	0.72	0.53	0.73	2.67	0.71	0.79	0.68	0.82	3.00	50.12
ReAct	0.68	0.75	0.65	0.77	2.85	0.68	0.70	0.52	0.71	2.61	0.70	0.75	0.66	0.80	2.91	49.55
RAD	0.69	0.76	0.66	0.79	2.90	0.69	0.71	0.53	0.72	2.65	0.71	0.77	0.67	0.81	2.96	49.80
Reflexion	0.72	0.80	0.69	0.82	3.03	0.72	0.74	0.55	0.75	2.76	0.73	0.81	0.70	0.84	3.08	50.65
Standard FOT	0.68	0.93	0.66	0.77	3.04	0.74	0.81	**0.68**	0.79	3.02	0.76	0.94	0.77	0.85	3.32	51.45
FOT (indep. vot.)	0.67	0.87	0.68	0.82	3.04	0.73	0.79	0.62	0.78	2.92	0.75	0.89	0.76	0.84	3.24	51.80
VOCOT	0.63	0.73	0.58	0.67	2.61	0.67	0.83	0.60	0.82	2.93	0.68	0.77	0.79	0.80	3.04	48.95
FOT (no VOCOT)	0.75	0.84	0.70	0.79	3.08	0.76	0.80	0.64	0.80	3.00	0.77	0.88	0.81	0.85	3.31	52.88
FOCOT (Ours)	**0.87**	**0.94**	**0.75**	**0.85**	**3.41**	**0.83**	**0.85**	**0.68**	**0.81**	**3.17**	**0.86**	**0.94**	**0.92**	**0.91**	**3.63**	**57.65**

gories. **Full descriptions for each baseline are available in the Appendix.**

The categories include: (i) **Naive baselines** (Random Guess, Majority Answer); (ii) **Standard reasoning paradigms** (Standard CoT (Wei et al., 2022), Self-Consistency CoT (Wang et al., 2023c), ToT (Yao et al., 2023b), ReAct (Yao et al., 2023c), RAD (Wang et al., 2022), and Reflexion (Shinn et al., 2023)); (iii) **FOT variants** (Standard FOT, FOT w/ indep. voting); (iv) **Grounding methods** (VOCOT (Wang et al., 2023b)).

Crucially, we also include two key ablations of our own framework:

- **FOCOT (no VOCOT)**: Our framework with node collaboration and contrastive decoding, but *without* spatial grounding.
- **FOCOT (Ours)**: The full model, integrating all three components.

This set of baselines provides a broad and rigorous comparison.

4.4 Main Results

We validated FOCOT against a comprehensive set of reasoning baselines on three key benchmarks: the multimodal reasoning benchmark **GQA** (Hudson and Manning, 2019), the hallucination-focused **POPE** (Lin et al., 2014) benchmark, and the **AMBER** (Wang et al., 2023a) dataset. The results on 7B backbones and the larger **DeepSeek-R1-Distill-Qwen-14B** (DeepSeek-AI et al., 2025) are presented in Table 1 and Table 2, respectively.

Our framework demonstrates state-of-the-art performance by effectively integrating its two core mechanisms. On **GQA**, FOCOT achieves a **3.30** (Volcano-7B) and **3.31** (LLaVA-1.6-7B), a significant lead over the strongest *Standard FOT* baselines. This gain is driven by our *grounded node-level collaboration* (Sec. 3.5), which aggregates supporting evidence from *partially correct* nodes to create a logically complete reasoning structure. On the **POPE** benchmark, FOCOT's **54.0%** accuracy achieves a massive +6.1% absolute gain over

Standard FOT. This result is directly attributable to our *node-level contrastive decoding (NCD)* (Sec. 3.6), which provides an active defense by semantically filtering visually-unsupported paths.

As shown in Table 2, FOCOT exhibits superior scalability when deployed on the **DeepSeek-R1-Distill-Qwen-14B** backbone. While the distilled R1 model provides a stronger reasoning baseline (e.g., Standard FOT achieving 3.04 on GQA), FOCOT further pushes the performance ceiling to **3.46** on GQA and **57.65%** on POPE.

Notably, although **VOCOT** excels in visual perception (achieving a top **0.84** in AMBER Vis.), it suffers from lower overall reasoning scores due to its fragile single-chain structure. In contrast, FOCOT successfully harnesses VOCOT's perceptual depth while mitigating its error propagation through our parallel forest architecture. This synergy allows FOCOT to outperform the *FOT (no VOCOT)* variant by **+0.38** on GQA Total and **+4.77%** on POPE accuracy, confirming that our framework remains highly effective as model capacity scales.

4.5 Ablation Studies

We evaluate the necessity of each component (NC, CD, VOCOT) and structural configurations (K, S), referencing Table 3b and Fig. 3.

Component Ablations. Starting from the full FOCOT (**GQA**=3.30, **POPE**=54.0%), we systematically ablate each mechanism. First, removing **Node-level Contrastive Decoding (CD)** drops POPE performance to 51.25%, confirming its role as the primary defense against hallucinations via token-level suppression. Second, removing **Grounded Node-level Collaboration (NC)** leads to a significant hit on GQA and AMBER, proving that cross-branch evidence aggregation is vital for logical coherence. Finally, the **VOCOT-only** variant (GQA=2.43) is insufficient alone. The full system significantly outperforms pairwise combinations (e.g., NC+VOCOT at 2.80 vs. 3.30), proving our components are *synergistic*: VOCOT provides grounded candidates, NC integrates evidence, and CD filters reasoning noise.

Structural and Efficiency Analysis To discern architectural innovation from simple ensemble effects, we compare FOCOT against two baselines. **Single-Path + CD** ($K = 1, S = 1$) yields the lowest GQA (2.05), underscoring that contrastive decoding requires a multi-branch search space to unlock its potential. Furthermore, under an equal budget (9 paths), **SC-CoT** (2.91) improves robustness via majority voting but remains significantly inferior to **Full FOCOT** (3.30). Unlike SC-CoT's independent paths, FOCOTs tree-structured collaboration enables iterative refinement and superior hallucination suppression. The performance gap between SC-CoT and FOCOT further confirms that while prompts provide diversity, our architecture is necessary to synthesize it into reliable answers.

Search-Budget Ablations. We analyze reasoning breadth (K) and depth (S) in Fig. 3. Scaling K or S from 1 to 3 yields substantial gains, while moving to 4 provides minimal quality improvement with increased latency. We quantify this via the **Efficiency Index**:

$$\mathcal{I}_{eff}(K,S) = \frac{E(K,S)}{\max_{k',s'} E(k',s')}, \quad E(K,S) = \frac{\mathcal{P}(K,S)}{\mathcal{T}(K,S)} \tag{8}$$

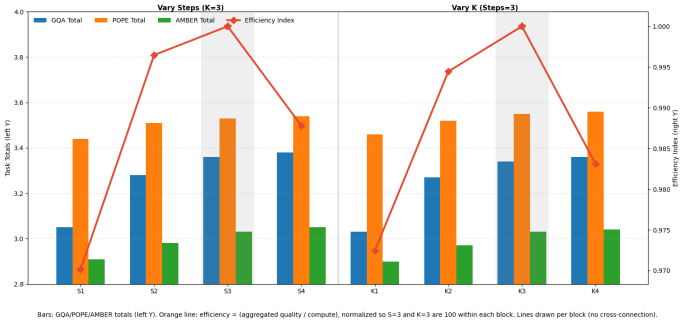
where $\mathcal{P}$ is the aggregate score and $\mathcal{T}$ is GPU-time. Efficiency peaks at $K = 3, S = 3$; scaling further results in diminishing returns and a sharp drop in $\mathcal{I}_{eff}$, validating our configuration.

Analysis by Dataset. FOCOT achieves superior precision on **GQA**, adversarial stability on **POPE** via CD-based hallucination suppression, and higher logical coherence on **AMBER** through well-ordered, evidence-backed explanations.

4.6 Qualitative Analysis

Figure 4 provides qualitative comparisons between the baseline and our **FOCOT** framework. Baseline models (orange boxes, "Without enhanced FOT") typically produce correct but minimal answers lacking sufficient justification. For instance, given the query *"Is it overcast?"*, the baseline simply outputs *"No. It is clear,"* a response that lacks explicit visual anchoring.

In contrast, the FOCOT framework (blue boxes, "With enhanced FOT") demonstrates a more robust and multi-faceted reasoning process. As illustrated in the "overcast" example, FOCOT actively identifies key visual evidence *("The sky is blue and clear")* before forming a logical conclusion. For the *"pillow on the couch"* task, while the baseline performs a simple verification, FOCOT employs a comprehensive decomposition: it first identifies all relevant entities *("There is a pillow and a couch")* and then analyzes their spatial relationship *("The pillow is on top of the couch")*. Notably, this reasoning is tightly coupled with **coordinate-level grounding** (blue box, "With coordinate"),



(a) Search budget ablation (varying $K = 1..4$ and $Steps = 1..4$). Bars show task totals (left Y), while the red line shows the **Efficiency Index** (right Y), which peaks at K=3 and $Steps$=3.

(b) **Component and Structural Analysis.** FOCOT vs. various baselines.

Method / Variant	NC	CD	VO	K	S	GQA Total	POPE Acc.	POPE Total	AMB. Total
Component Ablations									
NC-only	✓			3	3	2.20	40.21	2.50	2.30
CD-only		✓		3	3	2.25	44.32	2.52	2.35
VOCOT-only			✓	3	3	2.43	45.43	2.84	2.74
NC + CD	✓	✓		3	3	2.55	50.10	2.80	2.55
NC + VOCOT	✓		✓	3	3	2.80	51.25	2.95	2.70
CD + VOCOT		✓	✓	3	3	2.82	51.72	2.93	2.71
Structural Baselines									
Single-Path				1	1	2.17	44.20	2.33	2.06
Single-Path + CD		✓		1	1	2.19	45.31	2.45	2.14
SC-CoT				9	1	2.91	50.45	3.08	2.69
FOCOT (Full)	✓	✓	✓	3	3	**3.30**	**54.00**	**3.51**	**3.00**

Figure 3: Ablation studies for components and search budget. **(a)** The search budget chart, where the Efficiency Index (red line) peaks at K=3, $Steps$=3. **(b)** The component ablation table, showing all three mechanisms are complementary.

where small image patches represent the precise spatial anchors derived from our coordinate tokens (§3.3). This ability to synthesize multi-perspective evidence through a structured forest ensures that FOCOT's reasoning is not only reliable but also fully interpretable, effectively overcoming the brittleness of single-path approaches.

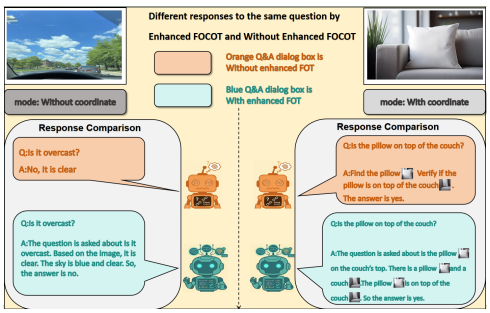


Figure 4: Qualitative examples from GQA showing evaluator feedback on baseline vs. FOT-enhanced answers. FOT outputs achieve higher semantic alignment and more comprehensive reasoning.

5 Conclusion

We presented **FOCOT**, a training-free framework that adapts Forest-of-Thought (FOT) (Bi et al., 2024) for complex multimodal reasoning via two synergistic mechanisms: *grounded node-level collaboration* (NC) and *node-level contrastive decoding* (NCD). FOCOT enhances reasoning reliability by: (a) aggregating complementary evidence through a collaborative graph; (b) actively suppressing hallucinations via semantic-level contrastive decoding; and (c) anchoring intermediate steps to precise image regions via coordinate tokens. Extensive evaluations on **GQA**, **POPE**, and **AMBER** demonstrate that FOCOT consistently achieves state-of-the-art performance, showing significant gains in accuracy and robustness across various backbones, including the **DeepSeek-R1-Distill-Qwen-14B**. As a model-agnostic, plug-and-play architecture, FOCOT provides a practical blueprint for enhancing large-scale multimodal systems without costly retraining.

Limitations

Despite the performance gains achieved by FOCOT, several limitations remain to be addressed in future work.

First, as illustrated in **Figure 6b**, FOCOT expands the reasoning space from a linear chain to a multi-branch tree to ensure robustness. However, this **structural complexity** inherently increases computational overhead. While multi-branch exploration (§3.4) effectively recovers from early visual misinterpretations, the necessity of maintaining and scoring multiple parallel paths (e.g., the "medal", "celebration", and "photo context" branches in Figure 6b) leads to higher inference latency compared to single-path decoding. This trade-off between reasoning breadth and real-time efficiency remains a challenge for deployment on edge devices.

Second, the **semantic aggregation** in FOCOT relies on the distinctiveness of the generated nodes. If the initial prompts fail to elicit sufficiently diverse reasoning trajectories (as shown in the bottom half of Figure 6b), the Node-level Contrastive Decoding (NCD) may lack a sufficiently "weak" context to contrast against, potentially diminishing its ability to suppress subtle hallucinations.

Third, FOCOT is a **training-free wrapper**. While this ensures high generalizability across backbones (as shown in Table 1), the model's ultimate performance is still upper-bounded by the base vision-language model's intrinsic perception. If the vision encoder fails to detect the core objects (e.g., the medal coordinates in Fig. 6b) entirely, FOCOT's collaborative mechanisms can only mitigate, but not fully rectify, such fundamental perceptual gaps.

Ethical Statement

We prioritize ethical integrity in our research. All experiments were conducted using academic-standard benchmarks, including GQA (Hudson and Manning, 2019), POPE (Lin et al., 2014), and AMBER (Wang et al., 2023a), which consist of publicly available data and contain no personally identifiable information (PII).

From a broader impact perspective, the **FOCOT** framework is specifically designed to **mitigate hallucinations** in multimodal models. By suppressing factually inconsistent reasoning via *Node-Level Contrastive Decoding*, our method directly contributes to the development of more reliable and trustworthy AI systems. Regarding content safety, while our framework significantly improves logical consistency and visual grounding, it is designed to be a **modular reasoning enhancer**. As such, it should be deployed in conjunction with dedicated safety-alignment filters and toxicity-detection protocols inherent to the respective backbone models to ensure responsible use in sensitive social contexts.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michał Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. 2024. Forest-of-thought: Scaling test-time compute for enhancing llm reasoning.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023b. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research (TMLR)*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. DoLa: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang Educational, Xingkai Yu, Yu Wu, Z.F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 3 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language

models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202. PMLR.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2024. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 406–425, Mexico City, Mexico. Association for Computational Linguistics.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Tsung-Yi Lin, Genevieve Patterson, Matteo R. Ronchi, Yin Cui, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Larry Zitnick, and Piotr Dollár. 2014. Coco:common objects in context. https://cocodataset.org/. Accessed: 2025-10-03.

Haotian Liu, Chunyuan Li, Qingyang Lin, Yong Jae Li, Zudi Hu, and Pengchuan Zhang. 2024. Llava: Large language and vision assistant. https://github.com/haotian-liu/LLaVA. GitHub repository.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. 2025. Argus: Vision-centric reasoning with grounded chain-of-thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual CoT: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *arXiv preprint arXiv:2403.16999*.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik R. Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2023a. AMBER: An automated multi-dimensional benchmark for multi-modal hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Xiaonan Wang and 1 others. 2023b. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models.

Xuezhi Wang, Jason Weiand Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Rationale-augmented ensembles in language models. *arXiv preprint arXiv:2207.00747*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023b. Tree of Thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023c. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. STaR: Bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2203.14465*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations (ICLR)*.

Appendix

A Evaluation Methodology Details

As mentioned in the main paper, we employ an evaluator model to score responses along four dimensions for each benchmark. This section provides the complete list of dimensions.

Algorithm 2: Multimodal Forest-of-Thought with VOCOT, Node Collaboration, and Contrastive Decoding

Input: Image I, question Q, backbone $\mathcal{M}$, branches K, temperature τ, grounding bonus β, contrastive weight λ
Output: Answer $\hat{A}$ and reasoning trace
Preprocess: encode I to region embeddings; tokenize Q; pack batch B.
$B \leftarrow \text{SanitizeBatch}(B)$
$\mathcal{P} \leftarrow \emptyset$
for $k = 1$ **to** K **do**
  $p_k \leftarrow \text{SeedPrompt}(Q)$
  for $t = 1$ **to** T **do**
    $p_k \leftarrow p_k \oplus \text{AspectPrompt}(t)$
    $y \leftarrow \text{SafeGenerate}(\mathcal{M}, B, p_k, \tau)$
    if *y contains structured coordinates (e.g., <click>)* **then**
      mark y as grounded
    end
    $n \leftarrow \text{SelectSentence}(y)$
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{n\}$; $p_k \leftarrow p_k \oplus n$
  end
end
Score nodes (contrast + grounding).
foreach $n_i \in \mathcal{P}$ **do**
  $s_i^{\text{lik}} \leftarrow \text{ScoreLik}(n_i)$ $s_i^{\text{neg}} \leftarrow \max_{j \neq i} s_j^{\text{lik}}$
  $\mathcal{L}_i^{\text{ctr}} \leftarrow -s_i^{\text{lik}} + \max(0, s_i^{\text{neg}} - s_i^{\text{lik}})$
  $s_i \leftarrow s_i^{\text{lik}} - \lambda \mathcal{L}_i^{\text{ctr}}$
  $g_i \leftarrow \mathbb{I}[\text{grounded}(n_i)]$
  $s_i' \leftarrow s_i + \beta\, g_i$
end
$\alpha_i \leftarrow \text{softmax}(\{s_i'\})$, $\quad h_{\text{final}} \leftarrow \sum_i \alpha_i\, h(n_i)$
Compose reasoning & answer.
$p^\star \leftarrow \text{ComposeThoughts}(\mathcal{P}, \alpha)$
$y_{\text{stage1}} \leftarrow \text{SafeGenerate}(\mathcal{M}, B, p^\star, \tau)$
if *y_{stage1} contains coordinates* **then**
  $B' \leftarrow \text{VoCoTRefine}(B, y_{\text{stage1}})$
  $\hat{A} \leftarrow \text{SafeGenerate}(\mathcal{M}, B', \text{FinalPrompt}, 0)$
end
else
  $\hat{A} \leftarrow \text{ExtractAnswer}(y_{\text{stage1}})$
end
return $\hat{A},\ p^\star \cup \{\textit{Final Answer} = \hat{A}\}$

Algorithm 3: Node-Level Contrastive Decoding (NCD)

Input: Nodes $\{\nu_i\}$ with reasoning r_i, prediction p_i, base utilities $\{s_i\}$; model $\mathcal{M}$; hyperparams k, m, α, β
Output: Refined scores $\{s_i'\}$ and (optional) aggregation weights $\{\alpha_i\}$

Confidence for context selection
foreach ν_i **do**
  $C_i \leftarrow S_{\text{len}}(r_i) + B_{\text{pred}}(p_i)$
end

Build strong/weak contexts
$\mathcal{I}_{\text{top}} \leftarrow$ indices of top-k C_i; $\quad \mathcal{I}_{\text{bot}} \leftarrow$ indices of bottom-m C_i
$X_{\text{strong}} \leftarrow \text{concat}(\{r_j : j \in \mathcal{I}_{\text{top}}\})$
$X_{\text{weak}} \leftarrow \text{concat}(\{r_j : j \in \mathcal{I}_{\text{bot}}\})$

Contrastive re-scoring (teacher-forced log-probs)
foreach ν_i **do**
  $L_{\text{str}} \leftarrow \sum_{t=1}^{|r_i|} \log p_{\mathcal{M}}(r_{i,t} \mid X_{\text{strong}}, r_{i,<t})$
  $L_{\text{weak}} \leftarrow \sum_{t=1}^{|r_i|} \log p_{\mathcal{M}}(r_{i,t} \mid X_{\text{weak}}, r_{i,<t})$
  $\Delta_{\text{CD}}(r_i) \leftarrow (1+\alpha)\, L_{\text{str}} - \alpha\, L_{\text{weak}}$
  $s_i' \leftarrow s_i + \beta\, \Delta_{\text{CD}}(r_i)$
end

GQA Dimensions. To provide a quantitative assessment of the model's performance on the GQA dataset, we define four metrics, ensuring each score is normalized within the $[0, 1]$ range:

1. **Object Similarity (S_{obj})**: To account for synonymous labels (e.g., "lady" vs. "woman"), we define a "Soft-Accuracy" using the rectified cosine similarity to ensure the score remains non-negative:

$$S_{obj} = \frac{1}{|O_{gt}|} \sum_{o_i \in O_{gt}} \max_{p_j \in O_{pred}} \left(\max \left(0, \frac{\phi(o_i) \cdot \phi(p_j)}{\|\phi(o_i)\| \|\phi(p_j)\|} \right) \right) \tag{9}$$

   where $\phi(\cdot)$ is a pre-trained semantic embedding function.

2. **Logical Completeness (L_{comp})**: This measures the recall of necessary logical operations. As a ratio of set cardinalities, it naturally falls within $[0, 1]$:

$$L_{comp} = \frac{|\mathcal{P}_{pred} \cap \mathcal{P}_{gt}|}{|\mathcal{P}_{gt}|} \tag{10}$$

   where $\mathcal{P}_{gt}$ represents ground truth logical predicates and $\mathcal{P}_{pred}$ represents predicted steps.

3. **Image Description Coverage (C_{img})**: This quantifies the fidelity of visual evidence reflection. The indicator function $\mathbb{1}(\cdot)$ ensures the mean remains within $[0, 1]$:

$$C_{img} = \frac{1}{|\mathcal{V}_{key}|} \sum_{v \in \mathcal{V}_{key}} \mathbb{1}(v \in \text{Output}) \tag{11}$$

   where $\mathcal{V}_{key}$ is the set of essential visual attributes.

4. **Thinking Comprehensiveness (T_{comp})**: To normalize the structural complexity, we di-

Algorithm 4: Evaluation Loop with FOT / Multi-Aspect / Monte-Carlo Switches

Input: Args: `use_fot`, `use_multi_aspect`, `enable_grounding`, `use_monte_carlo`, temperatures, thresholds
Output: Per-question predictions and logs
Load model and dataset; set image root; configure switches.
if `use_fot` **then**
  if `use_multi_aspect` **then**
    mode ← Multi-Aspect FOT (+grounding if enabled)
  else
    if `use_monte_carlo` **then**
      mode ← Monte-Carlo FOT
    else
      mode ← Standard FOT
    end
  end
else
  mode ← Non-FOT baseline
end
foreach *question* (I, Q) **do**
  run Alg. 2 with current mode; save $\hat{A}$ and trace
end
return results (JSON) and configs

vide the depth and breadth by their respective maximum observed values in the dataset, D_{max} and B_{max}:

$$T_{comp} = \alpha \cdot \frac{\text{Depth}(\mathcal{G}_r)}{D_{max}} + (1-\alpha) \cdot \frac{\text{Breadth}(\mathcal{G}_r)}{B_{max}} \tag{12}$$

where $\mathcal{G}_r$ is the reasoning DAG, and $\alpha \in [0, 1]$ is a weighting hyperparameter.

Scoring Integration. For the Amber and POPE benchmarks, we apply similar normalization principles. Specifically, to maintain the strictly $[0, 1]$ range while rewarding models for coordinate grounding, we define the final adjusted score S_{final} as a weighted sum:

$$S_{final} = 0.9 \cdot S_{base} + 0.1 \cdot \mathbb{1}(\text{valid coordinate grounding}) \tag{13}$$

where S_{base} is the normalized score derived from the dimensions above. This prevents the total score from exceeding 1.0 while explicitly incentivizing spatial reasoning.

POPE Dimensions. In addition to reporting standard performance, we employ four evaluator dimensions tailored to assess robustness and interpretability under the Probing Object Hallucination Evaluation (POPE) framework, with all scores normalized to the $[0, 1]$ range:

1. **Accuracy Consistency (A_{cons})**: This metric represents the "true accuracy" by requiring stability across perturbations. It naturally falls within $[0, 1]$ as an average of binary indicators:

$$A_{cons} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{1}(\text{pred}_i = y_i) \cdot \mathbb{1}(\text{pred}'_i = y_i) \right) \tag{14}$$

where pred_i and pred'_i are the model's responses to the original and perturbed queries, respectively.

2. **Reasoning Quality (Q_{re})**: We evaluate the soundness of the reasoning trace by computing the normalized similarity to a verified reference $\mathcal{R}_{ref}$:

$$Q_{re} = \frac{\text{score}(\text{Trace}_{pred}, \mathcal{R}_{ref})}{S_{max}} \tag{15}$$

where S_{max} is the maximum possible score attainable under the scoring heuristic, ensuring $Q_{re} \in [0, 1]$.

3. **Visual Understanding (V_{und})**: This assesses the grounding of visual attributes. To maintain the $[0, 1]$ range, we apply the rectified cosine similarity to the attribute embeddings ψ:

$$V_{und} = \frac{1}{|\mathcal{A}_{vis}|} \sum_{a \in \mathcal{A}_{vis}} \max\left(0, \frac{\psi(a_{pred}) \cdot \psi(a_{gt})}{\|\psi(a_{pred})\| \|\psi(a_{gt})\|}\right) \tag{16}$$

4. **Response Clarity (C_{res})**: We measure the lack of ambiguity using normalized Shannon entropy. By dividing the entropy by the maximum possible entropy ($\log |V|$), we ensure the final clarity index is strictly between 0 and 1:

$$C_{res} = 1 - \left(\frac{-\sum_{p \in P(y|x)} p \log p}{\log |V|} \right) \tag{17}$$

where V is the vocabulary size (or the set of candidate answer tokens). A value of 1 indicates total certainty, while 0 indicates maximum ambiguity.

Amber Dimensions. Since Amber focuses on open-ended generation without fixed ground truth, we formalize four quantitative metrics to evaluate reasoning quality, ensuring each is strictly normalized within the $[0, 1]$ range:

1. **Reasoning Logic** (L_{re}): This measures the structural coherence of the reasoning chain $\mathcal{C}$. Defined as the average transition probability between reasoning states s_t, it naturally falls within $[0, 1]$:

$$L_{re} = \frac{1}{|\mathcal{C}| - 1} \sum_{t=1}^{|\mathcal{C}|-1} P(s_{t+1}|s_t, I) \quad (18)$$

   where I represents the visual context.

2. **Visual Understanding** (V_{und}): This captures the accuracy of attribute interpretation and spatial grounding. It is calculated via the Jaccard similarity between predicted visual entities (including grounded coordinates) E_{pred} and verifiable scene attributes $\mathcal{V}$:

$$V_{und} = \frac{|E_{pred} \cap \mathcal{V}|}{|E_{pred} \cup \mathcal{V}|} \quad (19)$$

3. **Knowledge Application** (K_{app}): This assesses the integration of world knowledge. We use an exponential decay function of the minimum distance d to a verified knowledge graph $\mathcal{G}_{world}$, mapping the score to $(0, 1]$:

$$K_{app} = \exp(-\min d(\text{claim}, \mathcal{G}_{world})) \quad (20)$$

4. **Response Completeness** (C_{comp}): This evaluates the coverage of essential task facets $\mathcal{Q}$. We use rectified cosine similarity to ensure the score remains within $[0, 1]$:

$$C_{comp} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \max(0, \text{sim}(\text{Output}, q)) \quad (21)$$

To ensure objectivity, all dimensions are scored quantitatively. For Object Similarity (GQA) and Accuracy Consistency (POPE), the scores are derived directly from the normalized accuracy. We adopt this "Soft-Accuracy" approach for Object Similarity to account for semantically equivalent but textually distinct answers (e.g., "lady" vs. "woman") produced by models like FOCOT and VOCOT. The final score S_{final} for each benchmark is the unweighted average of its respective dimensions, ensuring a consistent evaluation scale across all models.

B Baseline Descriptions

We compare our method against a wide range of baselines, as summarized in the main paper. The full descriptions are as follows:

- **Random Guess**: A lower bound that randomly selects answers without any reasoning.
- **Majority Answer**: A heuristic baseline that always outputs the most frequent label (e.g. "yes" or common objects).
- **Vanilla LLM (no CoT)**: Direct generation from the multimodal backbone without chain-of-thought prompting.
- **Standard CoT**: Single-path chain-of-thought prompting, representing the most common reasoning strategy (Wei et al., 2022).
- **Self-Consistency CoT**: Multi-sample CoT reasoning where multiple paths are generated and a majority vote is taken (Wang et al., 2023c).
- **Tree-of-Thought (ToT)**: Reasoning as a search process over a tree of intermediate thoughts, guided by scoring and pruning strategies (Yao et al., 2023b).
- **ReAct**: An interleaved reasoning-and-acting framework that combines CoT-style reasoning with environment interaction (Yao et al., 2023c).
- **Rationale-Augmented Decoding (RAD)**: Enhances CoT reasoning by sampling diverse rationales and selecting the most consistent prediction (Wang et al., 2022).
- **Reflexion**: A self-improvement framework where the model reflects on its own reasoning trajectories and iteratively refines them (Shinn et al., 2023).
- **Standard FOT**: Multi-path reasoning with independent branches, without collaboration or grounding.
- **FOT (indep. voting)**: Similar to Standard FOT, but final predictions are aggregated via majority voting across branches.

- **VOCOT**: Chain-of-thought reasoning enhanced with coordinate grounding (Wang et al., 2023b), without multi-branch exploration.
- **FOCOT (no VOCOT)**: Our framework with node collaboration and contrastive decoding, but without spatial grounding.
- **FOCOT (Ours)**: The full model, integrating node collaboration, contrastive decoding, and VOCOT grounding.

C Statement on the Use of AI Assistants

In accordance with the ACL Rolling Review (ARR) policy regarding the use of AI assistants in scientific writing, the authors declare the following:

- **Linguistic Optimization:** We utilized the Gemini 2.5 Pro model specifically to assist in linguistic polishing, grammar correction, and improving the stylistic flow of the manuscript.
- **Originality of Content:** The initial draft, including all core research ideas, the design of the FOCOT framework (specifically Node-level Contrastive Decoding and Grounded Node-level Collaboration), and the experimental analysis, was entirely conceived and written by the authors.
- **Verification:** All AI-generated suggestions were manually reviewed, verified, and refined by the authors to ensure technical accuracy and to maintain the original research intent.
- **Experimental Role:** As explicitly detailed in Section 4.2, Qwen-turbo was employed as a large-scale automated evaluator for model-based scoring; this application is distinct from the writing assistance described above and constitutes a core part of our experimental methodology.

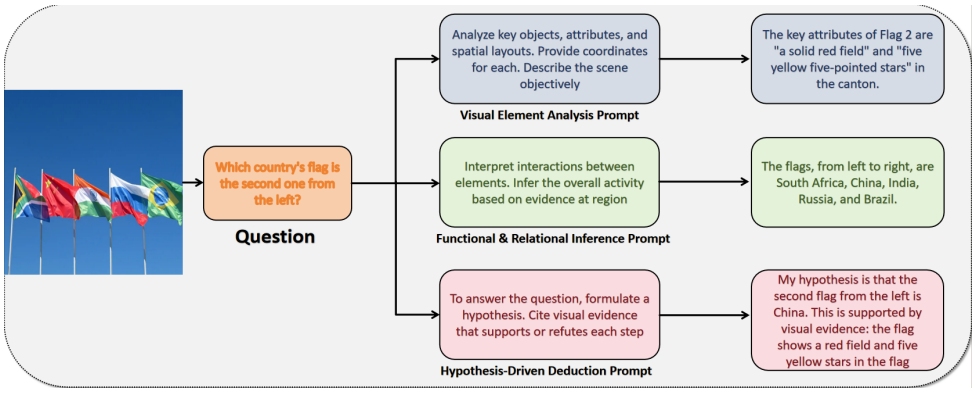
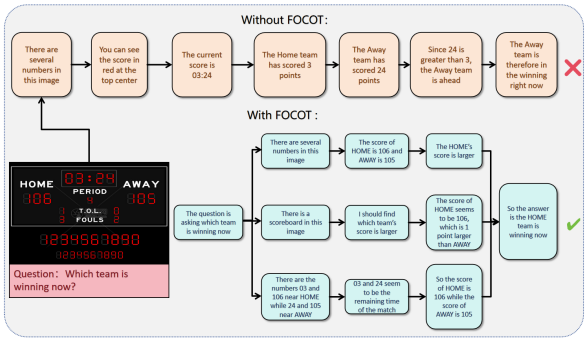
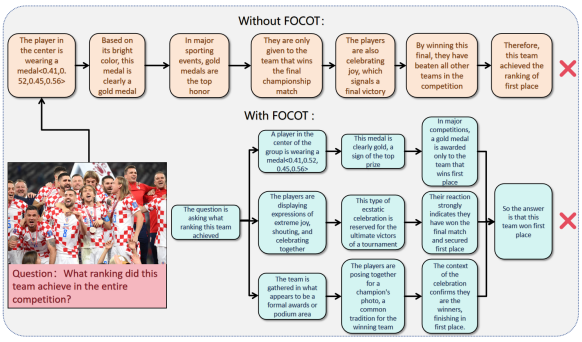


Figure 5: Visualization of our guided prompt mechanism. The input **Question** (orange) is fed into three parallel prompt-guided generators: (1) The **Visual Element Analysis Prompt** (blue) elicits objective descriptions of objects and attributes. (2) The **Functional & Relational Inference Prompt** (green) encourages the model to infer interactions and the overall scene context. (3) The **Hypothesis-Driven Deduction Prompt** (red) forces the model to formulate and test a hypothesis, citing specific evidence. These three distinct outputs then form the initial set of nodes in our Forest of Thoughts.



(a) A **success case** where FOCOT's multi-branch reasoning correctly identifies the winning team by analyzing multiple pieces of evidence (score, team names) on a complex, reflective scoreboard, while the baseline fails.



(b) A **failure case (limitation)** where FOCOT misidentifies the team's ranking. Despite multi-path reasoning, all branches are misled by strong visual glare and reflection on the medals, leading to an incorrect conclusion.

Figure 6: Qualitative examples of FOCOT's performance. **(a)** A success case demonstrating robust reasoning on a complex scoreboard. **(b)** A failure case illustrating a limitation where FOCOT is misled by challenging visual conditions (e.g., reflections).