Supplementary Materials for the following paper: Multimodal Graph-of-Thoughts: Hypothesis-Verification Graphs for Multimodal Reasoning in Vision-Language Models

Contents

Preface

In this supplementary material, we provide additional detail to enhance the main paper's content.

A. Cross-Modal Verification Functions

This section provides the exact definitions of the three cross-modal verification functions used by MM-GoT in the main paper. These functions are the concrete verifier components of the verification-aware node score in Eq. (3) of the main paper. Each verifier takes as input a candidate textual hypothesis h_i at node v_i together with the input image $\mathcal{I}$, and returns a scalar score in $[0, 1]$, where larger values indicate stronger cross-modal support.

A.1. VQA-based Semantic Consistency

The semantic consistency verifier, denoted ϕ_{VQA}, measures whether the reasoning step expressed in h_i is semantically supported by the visual evidence in $\mathcal{I}$. Given node v_i, we query a frozen VQA model with the prompt *"Is the following reasoning step correct?"*, using h_i as the candidate reasoning step and conditioning on the image. We define

$$\phi_{\mathrm{VQA}}(v_i) = P_{\mathrm{VQA}}(\text{“Yes”} \mid h_i, \mathcal{I}) \in [0, 1], \tag{1}$$

where P_{VQA} denotes the probability assigned by the frozen VQA model to an affirmative response. Higher values of $\phi_{\mathrm{VQA}}(v_i)$ indicate that the proposed reasoning step is more semantically consistent with the image.

During search, the semantic verifier is activated only when

$$\phi_{\mathrm{VQA}}(v_i) > \tau_{\mathrm{VQA}}, \tag{2}$$

where $\tau_{\mathrm{VQA}} = 0.6$ is selected on the validation split.

A.2. Grounding-based Spatial Validity

The spatial validity verifier, denoted ϕ_{GND}, measures whether entities referenced in the hypothesis h_i can be localized in the image with high confidence and limited ambiguity. Let $\mathcal{E}(h_i)$ denote the set of visual entity mentions extracted from h_i. For each entity $e \in \mathcal{E}(h_i)$, we obtain a set of candidate detections from a frozen grounding model:

$$\mathcal{B}(e) = \{(b_k, s_k)\}_{k=1}^{K}, \tag{3}$$

where b_k is a predicted bounding box and $s_k \in [0, 1]$ is its confidence score. Let $s^{(1)}(e)$ and $s^{(2)}(e)$ denote the top-1 and top-2 detection scores for entity e, respectively. We define

$$\phi_{\mathrm{GND}}(v_i) = \frac{1}{|\mathcal{E}(h_i)|} \sum_{e \in \mathcal{E}(h_i)} s^{(1)}(e)\, \sigma\left(\frac{s^{(1)}(e) - s^{(2)}(e)}{\gamma} \right), \tag{4}$$

where $\sigma(\cdot)$ is the logistic function and γ controls the softness of the uniqueness margin. This formulation favors both confident localization and unambiguous grounding. When $\mathcal{E}(h_i) = \varnothing$, the grounding verifier is not applied.

During search, the spatial verifier is activated only when

$$\phi_{\mathrm{GND}}(v_i) > \tau_{\mathrm{GND}}, \tag{5}$$

where $\tau_{\mathrm{GND}} = 0.4$ is selected on the validation split.

A.3. Attention-based Grounding Consistency

The attentional grounding verifier, denoted ϕ_{ATT}, measures whether image regions attended to by the MLLM align with the regions expected from the hypothesis. Let $A \in \mathbb{R}^{N \times M}$ denote the cross-attention map from the final layer of the frozen MLLM when processing h_i with image $\mathcal{I}$, where N is the number of text tokens and M is the number of image patches. For entity-referring tokens in h_i, we compute an attention distribution a_{entity} over image patches. We compare it against an expected spatial prior a_{expected}, derived from prior grounded nodes or from available supervision when present, using cosine similarity:

$$\phi_{\text{ATT}}(v_i) = \cos(a_{\text{entity}}, a_{\text{expected}}) \in [0, 1]. \tag{6}$$

Higher values indicate that the model's internal attention is concentrated on image regions compatible with the hypothesized reasoning step.

During search, the attentional verifier is activated only when

$$\phi_{\text{ATT}}(v_i) > \tau_{\text{ATT}}, \tag{7}$$

where $\tau_{\text{ATT}} = 0.5$ is selected on the validation split.

Together, ϕ_{VQA}, ϕ_{GND}, and ϕ_{ATT} provide complementary semantic, spatial, and attention-level evidence for ranking candidate nodes during graph search. Their outputs are combined with the LM prior in the node score defined in Eq. (3) of the main paper. In the ablation study in Sec. 5.1, we enable or disable individual verifiers by zeroing the corresponding terms while keeping graph construction, search topology, and synthesis fixed.

B. Ablation Study

B.1. Cross-Benchmark Performance Profile

Evaluating a reasoning framework on a limited set of benchmarks risks tying conclusions to specific task formats or visual domains. To assess whether MM-GoT generalizes beyond any single benchmark, we evaluate all methods on four datasets spanning abductive and defeasible multimodal reasoning (BlackSwan), broad multimodal perception and cognition (MME), expert-level multimodal understanding (MMMU-Pro), and general multimodal reasoning (MMStar). Table S.1 and Figure S.2 summarize the resulting cross-benchmark performance profile. To reduce backbone-specific effects, we report results averaged across all evaluated MLLM backbones (Qwen3-VL-Thinking, DeepSeek-VL2, GLM-4.6V, and InternVL3), and provide per-backbone breakdowns in Supplementary Table S.2.

Several patterns emerge. First, MM-GoT achieves its largest gains on BlackSwan and MMStar, where reasoning requires resolving ambiguity and maintaining multiple competing hypotheses under perceptual evidence. Second, gains on MMMU-Pro remain consistent across backbones, indicating that MM-GoT improves expert-level multimodal understanding beyond any single model family. Third, improvements on MME are smaller in absolute terms but remain positive, which is expected on a broad perception-and-cognition benchmark with less headroom for structured search. Finally, the relative ordering of methods (CoT, MM-CoT, ToT, GoT, and MM-GoT) is preserved across all four benchmarks, suggesting that the progression from linear prompting to verification-constrained graph search reflects a stable hierarchy of inference-time reasoning capability rather than benchmark-specific tuning. Collectively, these results support MM-GoT as a general-purpose multimodal reasoning framework.

B.2. Ablating Pruning and Ranking Policies

To test the hypothesis that verification improves search efficiency by suppressing inconsistent branches early, we ablate pruning and ranking policies while holding the search expansion budget (maximum node expansions) fixed. We compare:

Supplementary Table S.1. Cross-benchmark performance profile averaged across all evaluated MLLM backbones (Qwen3-VL-Thinking, DeepSeek-VL2, GLM-4.6V, and InternVL3). ΔACC denotes the absolute gain in percentage points (p.p.) of MM-GoT over GoT. The corresponding plot is shown in Fig. S.2.

Benchmark	**ACC (%)**					**ΔACC**
	CoT	**MM-CoT**	**ToT**	**GoT**	**MM-GoT**	
BlackSwan	57.3	58.6	59.9	60.7	**65.5**	+4.8
MME	75.5	76.0	76.0	76.3	**77.3**	+0.9
MMMU-Pro	60.4	61.3	62.1	62.8	**66.2**	+3.4
MMStar	62.4	63.3	64.0	64.7	**68.1**	+3.4
Avg	64.0	64.8	65.5	66.1	**69.3**	**+3.1**

Avg is the unweighted mean across the four benchmarks.
ΔACC is computed relative to GoT: $\Delta\text{ACC} = \text{ACC}_{\text{MM-GoT}} - \text{ACC}_{\text{GoT}}$ (p.p.).

Supplementary Table S.2. Per-backbone breakdown of cross-benchmark performance, ordered by decreasing Full MM-GoT average accuracy. We report GoT, Full MM-GoT, and ΔACC (p.p.).

Benchmark	**ACC (%)**											
	GLM-4.6V			**Qwen3-VL-Thinking**			**InternVL3**			**DeepSeek-VL2**		
	GoT	**MM-GoT**	**ΔACC**	**GoT**	**MM-GoT**	**ΔACC**	**GoT**	**MM-GoT**	**ΔACC**	**GoT**	**MM-GoT**	**ΔACC**
BlackSwan	64.1	68.5	+4.4	62.7	68.3	+5.6	59.8	64.9	+5.1	56.1	60.4	+4.3
MME	78.9	79.8	+0.9	77.6	78.8	+1.2	75.4	76.3	+0.9	73.4	74.1	+0.7
MMMU-Pro	66.9	69.8	+2.9	63.8	67.5	+3.7	62.1	65.8	+3.7	58.3	61.5	+3.2
MMStar	68.4	71.0	+2.6	66.4	70.6	+4.2	63.8	67.4	+3.6	60.1	63.2	+3.1
Avg	69.6	72.3	**+2.7**	67.6	71.3	**+3.7**	65.3	68.6	**+3.3**	62.0	64.8	**+2.8**

Avg is the unweighted mean across the four benchmarks.
ΔACC denotes the absolute gain of MM-GoT over GoT in percentage points (p.p.).
Wins: Full MM-GoT outperforms GoT on **4/4** benchmarks for each backbone.

(i) No pruning. All candidate nodes are retained up to the expansion limit, maximizing branching and allowing inconsistent branches to persist throughout the search.

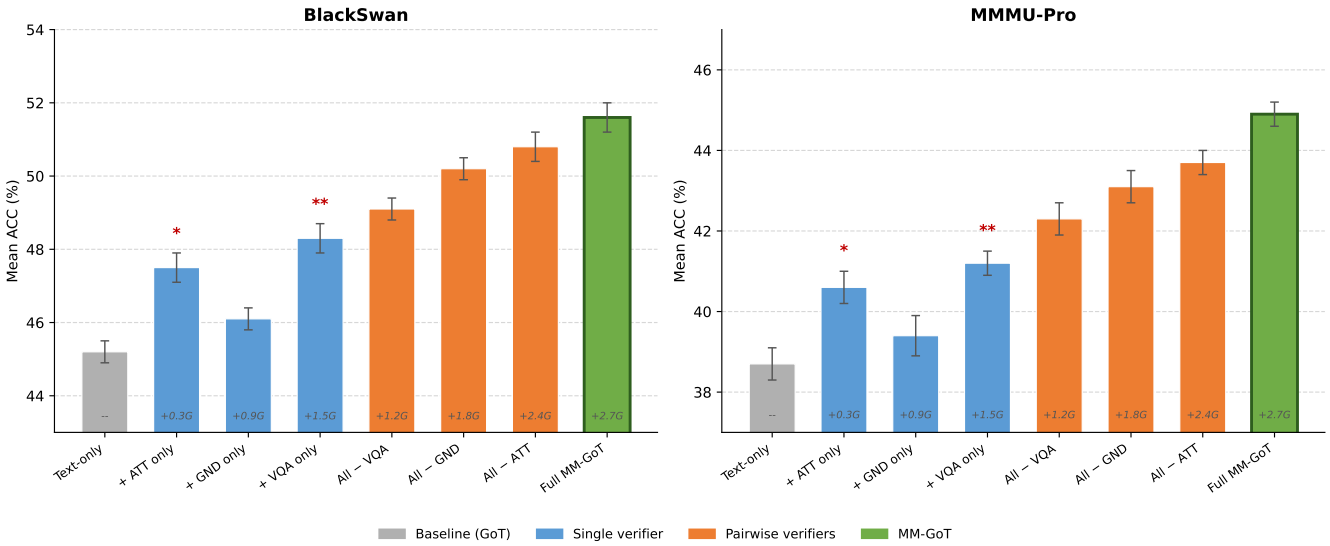
(ii) Likelihood-only pruning. Nodes are pruned using text-likelihood scores rather than verification scores, isolating the effect of pruning from the effect of cross-modal verification signals.

(iii) Verification-based pruning (MM-GoT default). Nodes are pruned using the verification-constrained objective in Eq. (5), suppressing visually inconsistent branches early in the search.

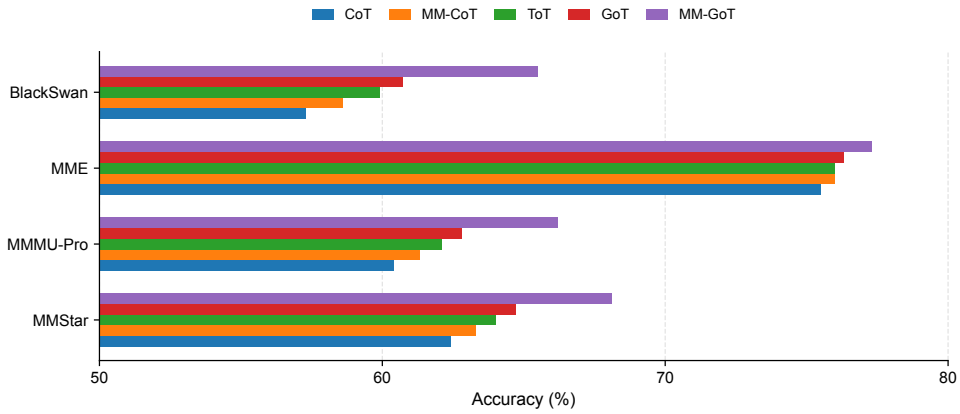
(iv) Threshold and top-k sweeps. We vary the pruning threshold τ and retained set size $k \in \{1, 2, 4, 8\}$ to characterize sensitivity and identify stable operating points across benchmarks.

Table S.3 reports mean accuracy (ACC (%)), tokens per query (Tok/Q), and the fraction of pruned nodes on BlackSwan and MMMU-Pro.[1] Verification-based pruning improves both accuracy and efficiency relative to likelihood-only pruning, indicating that cross-modal verification removes branches that are textually plausible yet visually inconsistent. Removing pruning entirely yields the highest token cost and the lowest accuracy, suggesting that retaining all branches amplifies perceptual errors rather than resolving them. Sweeps over τ show a broad stable region for $\tau \in [0.3, 0.6]$, with performance degrading under overly aggressive pruning ($\tau = 0.2$) and under-pruning ($\tau = 0.8$), motivating $\tau = 0.5$ as a robust default. Top-k sweeps indicate that $k = 4$ provides a favorable accuracy–efficiency trade-off, consistent with the branching-factor analysis in Sec. 5.2 of the main text.

[1] In our implementation, smaller τ increases pruning aggressiveness, while larger τ retains more candidates.



Supplementary Figure S.1. Per-benchmark accuracy breakdown for cross-modal verification ablations. Mean ACC (%) ± std over 64 runs on BlackSwan (left) and MMMU-Pro (right) for each verifier configuration. ΔFLOPs (G) annotations are shown below each bar. $^{*}p < 0.05$, $^{**}p < 0.01$ (paired bootstrap, Bonferroni-corrected). Colors correspond to Figure 3 of the main paper.



Supplementary Figure S.2. Cross-benchmark performance profile. Mean ACC (%) on BlackSwan, MME, MMMU-Pro, and MMStar for CoT, MM-CoT, ToT, GoT, and MM-GoT, averaged across evaluated MLLM backbones. Corresponding values are reported in Table S.1, with per-backbone results in Supplementary Table S.2.

B.3. Budget Sensitivity and Scaling

We study how MM-GoT scales with inference-time compute by varying the per-query ΔFLOP budget while holding the backbone, prompt template, and maximum reasoning depth fixed. All methods share the same candidate-generation procedure and expansion schedule at each budget level, so observed differences reflect compute utilization rather than decoding artifacts.

Sample efficiency. Fig. 3 (main text) shows that MM-GoT's accuracy gains are front-loaded within the compute budget. At a per-query overhead of only $0.3{\times}10^9$ ΔFLOPs (achieved by enabling the attention-consistency verifier alone), mean accuracy across BlackSwan and MMMU-Pro rises from 41.95% (text-only baseline) to 44.05%, a gain of $+2.1$ p.p. for roughly 11% of the full MM-GoT overhead. Adding pairwise verifiers further increases accuracy to 46.65% at $1.8{\times}10^9$ ΔFLOPs, recovering approximately 65% of the full-system gain at approximately 67% of its cost. Full MM-GoT reaches 48.25% ($+6.3$ p.p. over the text-only baseline) at 2.7×10^9 ΔFLOPs.

These gains arise because verification concentrates compute on visually consistent hypotheses and suppresses perceptually inconsistent branches early, preventing them from accumulating downstream token cost; by contrast, likelihood-driven search allocates compute more uniformly across candidates regardless of visual validity.

Diminishing returns at high budgets. The graph-structure and synthesis ablation in Fig. 4 (main text) illustrates the opposite end of the scaling curve. Increasing the branching factor from $k{=}4$ (MM-GoT) to $k{=}8$ yields only $+0.2$pp accuracy at $1.9\times$ the per-query ΔFLOP cost, placing $k{=}8$ off the accuracy–efficiency Pareto frontier. This saturation arises because, at high budgets, the residual expanded branches are already largely visually consistent; additional exploration refines plausible hypotheses rather than recovering from perceptual errors, yielding negligible accuracy improvement per unit of added compute.

Practical operating region. Together, the two Pareto frontiers define a favorable operating region between roughly 1.2–2.7×10^9 ΔFLOPs, where additional compute yields consistent accuracy gains and MM-GoT dominates the ablated variants in the accuracy–efficiency trade-off. Outside this range—either under extremely tight budgets where aggressive pruning reduces candidate diversity, or under very large budgets where verification becomes redundant—marginal returns diminish. Our default setting of $k{=}4$ lies near the knee of this curve and is used throughout the paper. Unlike scalar reward models employed in Best-of-N test-time scaling [1, 2], MM-GoT's verification signals are compositional and modality-grounded, enabling informative pruning at low budgets rather than relying on post-hoc candidate re-ranking.

Supplementary Table S.3. Pruning policy ablations on BlackSwan and MMMU-Pro. Mean ACC (%) ± std over 64 runs. Tok/Q is the average number of tokens per query. Pruned (%) is the fraction of candidate nodes removed before synthesis.

Configuration	**BlackSwan**	**MMMU-Pro**	**Tok/Q**	**Pruned (%)**
Pruning strategy				
No pruning	49.1±0.4	42.6±0.3	2840	0.0
Likelihood-only	50.3±0.3	43.4±0.4	2210	31.4
Verif.-based (MM-GoT, default)	**51.6**±0.4	**44.9**±0.3	1980	38.7
Threshold τ sweep (verif.-based, $k = 4$)				
$\tau = 0.2$ (over-pruning)	47.3±0.5	40.8±0.4	1340	61.2
$\tau = 0.3$	50.9±0.3	44.1±0.3	1820	42.3
$\tau = 0.5$ (default)	**51.6**±0.4	**44.9**±0.3	1980	38.7
$\tau = 0.6$	51.4±0.3	44.6±0.4	2050	35.1
$\tau = 0.8$ (under-pruning)	50.1±0.4	43.2±0.3	2560	18.3
Top-k retained nodes sweep ($\tau = 0.5$)				
$k = 1$	48.4±0.3	41.7±0.4	1420	52.1
$k = 2$	50.2±0.4	43.3±0.3	1710	44.6
$k = 4$ (default)	**51.6**±0.4	**44.9**±0.3	1980	38.7
$k = 8$	51.7±0.3	45.0±0.4	2640	24.3

References

[1] Aisha Khatun and Daniel G. Brown. A study on large language models' limitations in multiple-choice question answering. *arXiv preprint arXiv:2401.07955*, 2024. 6

[2] Charlie Snell, Kanishk Jaeckle, Aviral Kumar, and Sergey Levine. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. 6