Autonomous Multimodal Reasoning via Implicit Chain-of-Vision

Yiqiao Huang[1,*], He Qi[2,*], Zhaorun Chen[3], Haopeng Zhang[4], Hanchao Yu[5], Zhuokai Zhao[3,5,†]
[1]Harvard University, [2]George Mason University, [3]University of Chicago [4]University of Hawaii, [5]Meta AI

Abstract

While large vision-language models (LVLMs) have made significant progress in multimodal reasoning, they continue to struggle with complex tasks requiring multistep reasoning involving different visual cues across reasoning stages. Specifically, LVLMs have difficulty focusing on critical image regions, limiting their ability to solve challenging multimodal algorithmic problems. To address this limitation, we propose ***Implicit Chain-of-Vision*** *(***ICoV***), a fine-tuning framework that empowers LVLMs to autonomously generate implicit rationales directly from visual inputs, improving reasoning capabilities without external supervision. Specifically,* ICoV *utilizes a step-by-step, decoupled training-inference framework, allowing the models to effectively integrate structured logical reasoning with targeted attention on essential visual regions during question answering. Experimental results demonstrate that* ICoV *significantly improves the performance of LVLMs on complex multimodal tasks, outperforming both standard fine-tuning methods and existing chain-of-vision (CoV)-based decoding approaches.*

1. Introduction

Multimodal, and more specifically vision-language reasoning [32, 39, 52, 59] has become a critical frontier in artificial intelligence (AI), where models are expected to understand comprehensive information from visual and textual inputs to reason and achieve high-level problem-solving ability and handle complex visual question answering (VQA) tasks. With the rapid development of large vision language models (LVLMs) [5, 17, 18, 20, 37, 50, 54, 58, 61], multimodal reasoning has advanced significantly, with models now able to handle increasingly complex queries by integrating visual and language understanding [4, 10, 19, 22, 27, 31, 48, 51].

However, most existing LVLMs unfortunately only perform well on straightforward comprehension of image content, falling short in more complex scenarios that require deeper or multistep reasoning [15, 21]. Specifically, a recent benchmark study [15], which focuses on evaluating the abstraction, deduction, and generalization capabilities of models in solving complex vision-language puzzles, argues that the current LVLMs are performing no better than second graders when the questions require more demanding reasoning skills.

Meanwhile, research community has observed much more success in large language model (LLMs) reasoning [24, 38, 56], where the model conducts reasoning solely based on textual inputs. Progress in LLM reasoning has shown impressive capabilities in handling tasks across diverse domains including mathematical problem-solving [1, 53], multi-step reasoning [11, 14, 44], and common sense reasoning [60]. Moreover, such performance can be further improved with either post-training finetuning [11, 53] or train-free prompting techniques [3, 46, 49].

This performance gap highlights the unique challenges of vision-language reasoning. One hypothesis [23] suggests that, while textual reasoning in LLMs benefits from structured linguistic patterns and explicit logical cues, visual information in LVLMs often lacks such inherent structure and requires additional contextualization to integrate with language-based reasoning. This hypothesis is partially shown empirically, where the multimodal chain-of-thought (MCoT) [59] improves vision-language reasoning using a two-stage framework that separates rationale generation and answer inference. By doing so, LVLMs can perform better on more complex, reasoning-intense questions as they can combine better generated rationales based on visual information *before* providing the final answer.

Despite the initial success of the approaches [32, 39, 57, 59], multimodal reasoning capabilities still remains extremely underexplored, especially compared to text-only LLM reasoning. Applying LLM reasoning techniques to LVLMs remain challenging, with the difficulty arises largely from the inherent complexity and ambiguity in interpreting visual inputs, which lack the structured cues available in textual reasoning. Additionally, prompt engineer-

*These authors contributed equally to this work. Correspondence to hqiao688@gmail.com

†All experiments, data collection, and processing activities were conducted by the collaborating institutions. Meta was involved solely in an advisory role and no experiments, data collection or processing activities were conducted using Meta tools or within its IT environment.

ing techniques [59] often fail to consistently elevate LVLM performance in scenarios requiring nuanced or multistep visual interpretation, particularly when models struggle with instruction following or reasoning [42].

To address these limitations, we propose **Implicit Chain-of-Vision (ICoV)**. ICoV is designed to internalize reasoning capabilities within LVLMs without relying solely on external LLM support. Specifically, ICoV utilizes a separated train-inference framework, where VLMs are fine-tuned to generate intermediate rationales directly from visual content before deriving final answers. This approach enables VLMs to autonomously build logical connections within visual data, enhancing their ability to tackle reasoning-heavy tasks. Experimental results show that, after training with ICoV, the model internalizes the reasoning methods from the fine-tuning dataset and effectively improves QA accuracy. Our method not only advances current LVLM capabilities but also demonstrates potential for broader applications in scenarios demanding complex multimodal reasoning.

2. Related Work

2.1. Vision-Language Reasoning

Existing LVLMs can generally maintain accuracy on simple image-based questions by accurately understanding image content. However, their performance diminishes when facing more challenging algorithmic problems based on images [15, 21]. To address these complex problems, recent studies have explored methods that integrates Chain-of-Thought (CoT) [46] into multimodal reasoning, with [34, 35] showing significant improvements in enhancing reasoning and response accuracy through logical chains.

However, existing studies focus primarily on CoT fine-tuning for LVLMs to enhance their text-based reasoning abilities. In this paper, we introduce a novel algorithm design called the Implicit Chain of Vision (ICoV), which utilizes multimodal reasoning techniques and integrates procedural fine-tuning with CoT decomposition. By progressively reducing the number of subproblems in fine-tuning at each training stage, we effectively enhance the image understanding, multimodal reasoning, and problem-solving capabilities of off-the-shelf LVLMs.

2.2. Multimodal CoT

Existing research on multimodal CoT with visual data [12, 40, 59] has explored various methodologies to enhance the reasoning capabilities of LVLMs. These methodologies can be broadly categorized into three types of methods: enhancing answer selection, integrating external reasoning resources, and internalizing logical reasoning.

Enhancing answer selection. Initially, studies often employ multiple VLMs to generate answers which are then selected through mechanisms like voting by large language models (LLMs) [4, 43]. Although this approach improved answer accuracy, it does not significantly enhance the VLMs' intrinsic reasoning capabilities.

Integrating external reasoning resources. Other studies have focused on incorporating external resources such as knowledge graphs or processing multiple images simultaneously to improve VLM's ability to connect images with questions [2, 9, 12, 13, 33, 40, 59]. These methods benefit comprehension and contextual understanding, yet they offer limited enhancements of the CoT reasoning specifically tailored for complex image-based problem-solving tasks.

Internalizing logical reasoning. More recent studies have utilized VLM-LLM integrations where visual and question information processed by VLMs is passed to LLMs. Then LLMs generate the rationale or logical chain for the final answer [6, 10, 28, 36, 41, 45, 47, 55]. This strategy utilizes LLM-generated CoT reasoning but often relies on external LLMs during inference, thus limiting end-to-end processing and under-utilizing LVLM's potential for independent reasoning. To address these limitations, our work introduces an implicit CoT (ICoV) training strategy that internalizes LLM-assisted reasoning capabilities within the VLM's structure, substantially improving response accuracy. This strategy builds upon prior research in reasoning fine-tuning, which has explored the combination of multimodal reasoning and reasoning fine-tuning strategy.

3. Methodology

Figure 1 illustrates the overall pipeline of our proposed ICoV, which comprises four core components: 1) *Tree-based Question Decomposition Synthesis*, 2) *Chain-of-Vision Image Decomposition*, 3) *localized visual attention*, and 4) *Stepwise Fine-tuning for Implicit Reasoning*. Specifically, the key insight of our approach is to leverage the fact that LVLMs perform better on VQA queries where the associated images contain simpler visual cues and less noise [12]. Therefore, ICoV decomposes complex multimodal questions into manageable sub-questions, each associated with a distinct local image region, effectively building up to solve the original complex question.

3.1. Tree-based Question Decomposition Synthesis

Chain-of-Thought (CoT) methods have recently shown notable success in enhancing reasoning capabilities by decomposing complex questions into simpler sub-questions, both in textual reasoning [46] and visual question answering (VQA) [59]. However, these methods typically excel at

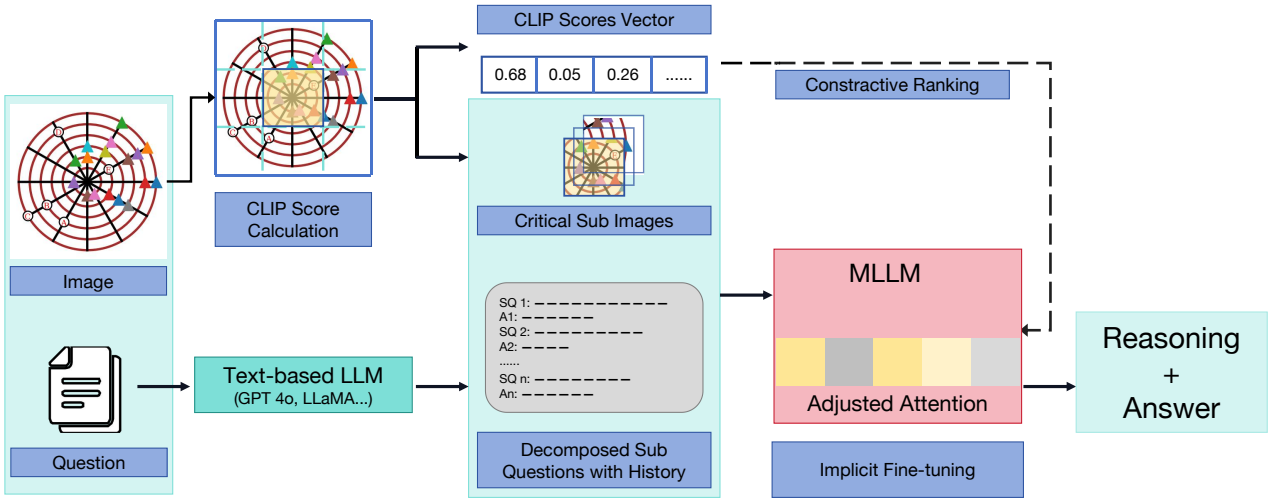

Figure 1. Overall architecture of the proposed pipeline. The pipeline begins with explicit chain-of-thought (CoT) prompting using GPT-4o to generate a sequence of sub-questions and corresponding answers (SQA) for each puzzle. For each SQA pair, CLIP-based similarity scores are computed to identify and rank the most relevant image regions, producing a structured sub-visual question-answer (SVQA) dataset. To ensure label quality, only SVQA branches leading to a correct final puzzle answer are retained. This curated dataset, comprising original puzzle questions, answers, and visual grounding from critical sub-images, is used to fine-tune a LVLM. During fine-tuning, we adjust the LVLM's cross-attention to focus on the identified sub-image regions, thereby enhancing its ability to reason over complex visual inputs.

general VQA tasks but struggle with decomposing intricate, multi-step reasoning questions (e.g., SMART-101 puzzle). Moreover, due to generalizability requirements, we assume no prior access to ground-truth solution paths, making existing CoT methods either inapplicable or inefficient, as valid solution paths are typically very sparse [11]. Additionally, current decomposition approaches rarely account for the inherently multimodal nature of visual reasoning, usually relying solely on textual decomposition.

To address this issue, we introduce an efficient tree-based multimodal question decomposition synthesis pipeline, as illustrated in Figure 2. To best exploit the fact that current MLLMs excel at reasoning over localized visual cues with minimal complexity (e.g., *counting objects*, *comparing lengths*), we first prompt a powerful MLLM (e.g., GPT-4o) to decompose the question into a sequence of sub-questions, each targeting a specific region of the image.

After generating the sub-questions, we employ a tree-based exploration method to derive ground-truth sub-answers. Specifically, (1) at each tree layer, we adopt multiple MLLMs to provide candidate answers for the current sub-question. Then, we use GPT-4o as a judge to select the most plausible solution paths up to that point [5]. Then, (2) we proceed to the next sub-question and extend each validated path by generating new sub-answer candidates, again using GPT-4o for selection. This stepwise expansion enables us to build a diverse set of accurate sub-solution paths. Finally, (3) upon reaching the leaf nodes (i.e the final sub-question), we retain only those paths whose final answers are verified to be correct. Notably, this synthesis pipeline enables the efficient collection of diverse yet accurate reasoning paths by leveraging both procedural and outcome supervision [11].

3.2. Chain-of-Vision Decomposition

While existing multimodal reasoning approaches mostly focus on textual decomposition, as discussed in Section 2, they often overlook the need for varying visual contexts at different reasoning steps. These methods typically provide the entire image for each sub-question, which can introduce unnecessary visual noise, particularly when the sub-question is relevant only to a specific region of the image.

To address this limitation, we propose *Chain-of-Vision (CoV) Decomposition*, which jointly decomposes both the textual question and the image, aligning each sub-question with a localized visual region to enable more focused and precise reasoning.

Specifically, we divide the original image into 13 sub-images designed to capture both granular and overlapping visual contexts. This includes a 3×3 grid layout that generates nine equal-sized segments, along with four additional sub-images centered at the intersections of the grid lines. Notably, all sub-images are maintained at the same resolution, ensuring consistency in patch tokenization and pre-

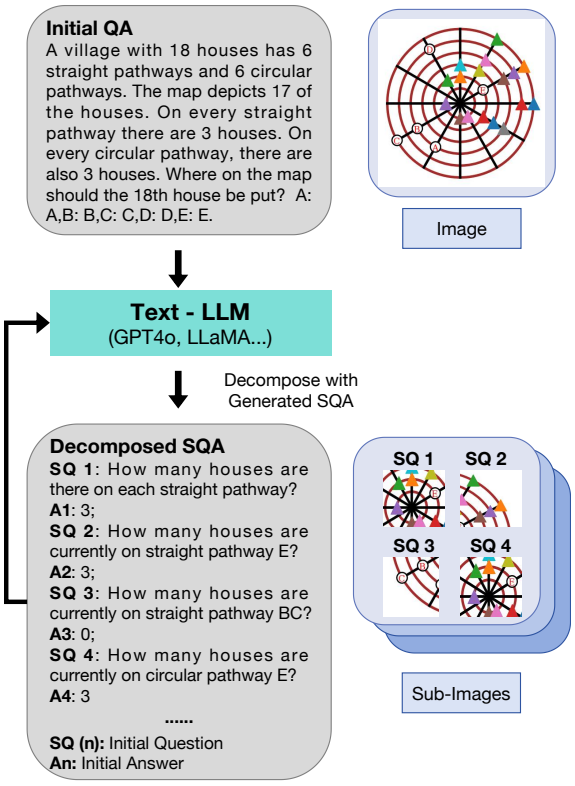

Figure 2. Multimodal decomposition framework. Given an input puzzle comprising both an image and a textual question, our method first performs textual decomposition by prompting an LLM to generate B root sub-question-answer (SQA) pairs. Each branch is recursively expanded using follow-up prompts until either the final answer is reached or the branch depth reaches D. Only branches that produce the correct final answer are retained for dataset inclusion. In parallel, we perform image decomposition by segmenting the original puzzle image into sub-images aligned with each SQA. These sub-images serve to visually ground the reasoning steps, enabling fine-grained multimodal alignment between sub-questions and relevant image regions.

serving spatial coherence. Specifically, this decomposition strategy allows the model to attend to distinct yet potentially overlapping regions of the image, which is particularly valuable for complex reasoning tasks that require interpreting multiple visual relationships in a structured sequence.

3.3. Localized Visual Attention

To incorporate this decomposition during fine-tuning, we pair each sub-question with the most relevant sub-image. The vision encoder processes these sub-images into patch tokens, which are then integrated into the model via cross-modal attention. We modify the cross-attention mechanism such that each text token selectively attends to the visual tokens of the associated sub-image, using a constrained attention mask to reinforce localized alignment.

To further enforce this targeted focus, we introduce an *attention-guided loss function*:

$$\mathcal{L} = -\sum_{i=1}^{N} \alpha_i \log(p_i) + \lambda |\mathbf{A} - \mathbf{A}^*|_2^2 \tag{1}$$

where N denotes the total number of image regions considered, α_i represents the attention weight assigned to the i-th region, and p_i is the predicted relevance probability for that region. The matrix A captures the model's actual attention distribution, while A^* denotes the ideal attention target derived from decomposition alignment. The term λ controls the balance between prediction accuracy and attention regularization.

By aligning sub-questions with localized image regions and guiding attention accordingly, our Chain-of-Vision decomposition significantly improves the model's ability to perform multi-step visual reasoning. This approach reduces ambiguity, enhances interpretability, and ensures more stable and accurate performance across complex vision-language tasks.

3.4. Stepwise Fine-Tuning for Implicit Reasoning

To enable LVLMs to perform implicit Chain-of-Vision (CoV) reasoning, we adopt a stepwise fine-tuning approach inspired by Stepwise Internalization [16]. This method incrementally transitions the model from relying on explicit supervision toward fully implicit multimodal reasoning. We begin with a base LVLM trained using the complete set of decomposed sub-question, sub-answer, and sub-image triplets, representing fully explicit CoV supervision. In each subsequent fine-tuning stage, we remove one sub-question-answer-image tuple from the beginning of the sequence for each puzzle, thereby reducing the model's exposure to intermediate reasoning steps.

Formally, at stage t, the model is fine-tuned using examples in which the first t reasoning steps are omitted. For our experiments, we set the removal step size $S = 1$, as it yields better empirical performance than larger step sizes. This process continues until only the original puzzle (comprising the original question, original image, and final answer) remains, effectively resulting in a model that relies solely on the implicit reasoning it has internalized during prior stages.

During inference, we evaluate the final-stage model without any intermediate reasoning inputs. The model is presented with original question and image and is expected to generate the correct answer based on its learned internal reasoning capabilities. This setup ensures zero reliance on external reasoning prompts, enabling efficient and scalable deployment in real-world multimodal reasoning tasks.

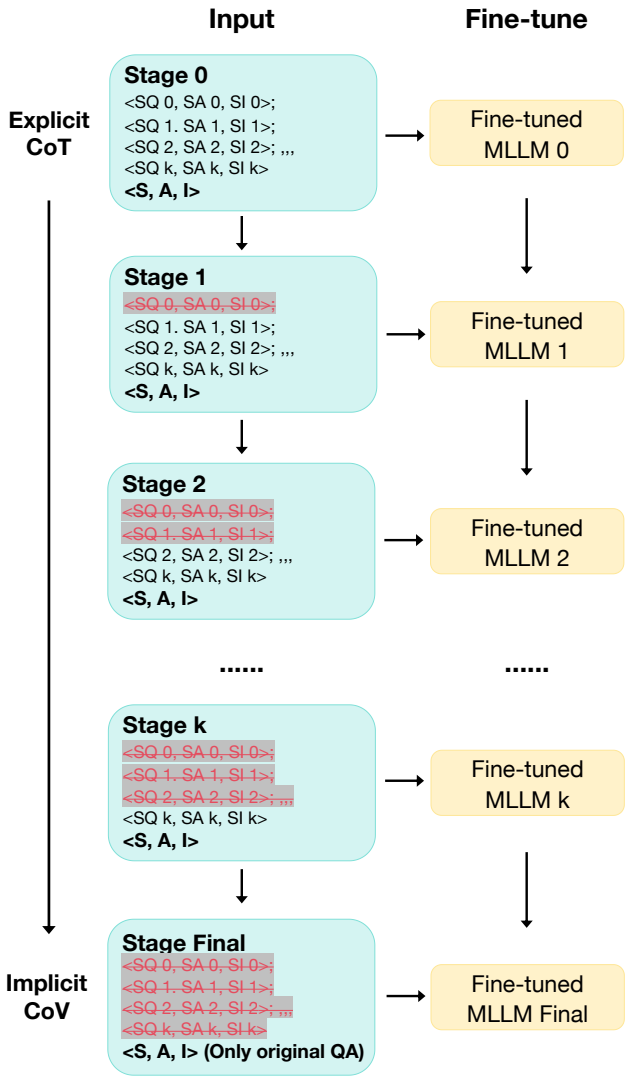

Figure 3. **Stepwise Internalization for ICoV.** The training process of ICoV's stepwise internalization method proceeds through a sequence of fine-tuning stages. At Stage 0, the model is trained with the full sequence of visual reasoning steps—represented as tuples $< \mathrm{SQ}_i, \mathrm{SA}_i, \mathrm{SI}_i >$ denoting sub-question, sub-answer, and sub-image, corresponding to explicit Chain-of-Vision (CoV) supervision. At each subsequent stage, one CoV token (from the beginning of the sequence) is removed, and the model is further fine-tuned to predict the final puzzle output using progressively less intermediate supervision. By the final stage, all intermediate CoV tokens have been removed, and the model is trained to solve the puzzle using only the original question-answer pair and image, representing implicit CoV reasoning.

4. Experiment

4.1. Experimental Setup

Fine-tuning dataset. We primarily use the SMART-101 dataset [15], which contains 101 unique puzzles requiring skills such as ordering, algebra, and spatial reasoning. Each puzzle includes 2,000 sub-puzzles. To enable ICoV reasoning in MLLMs, we construct a decomposition dataset by combining text-based and vision-based reasoning optimization strategies, as detailed in Section 3.

For each puzzle, we decompose it into $B = 3$ main branches or *subroot puzzles*, each containing up to $S = 10$ follow-up sub-questions, resulting in 60,000 total sub-questions and sub-answers per puzzle. We select nine representative puzzle types (IDs: 18, 61, 62, 69, 71, 73, 77, 94, and 99) based on the **Instance Split** outlined in the original SMART-101 paper [15].

To fine-tune the MLLMs, we include only sub-questions and sub-answers where GPT-4o predicted a correct final answer during the decomposition process. If multiple branches yielded the correct final answers, we retain only the first correct branch in the data set. Details of the quality of decomposition for each puzzle are summarized in Table 1. Due to the high cost of preparing this dataset, we focus our experiments on three particularly challenging puzzles – Puzzle 18, 69, and 99 – where even state-of-the-art models like GPT-4o exhibit notably low accuracies.

Puzzle ID	Sub-Question Level			Branch Level		
	Q_F	Q_C	Q_{acc} (%)	B_F	B_C	B_{acc} (%)
18	1,683	542	32.20	3,131	683	21.81
69	1,996	838	41.98	5,541	1,114	20.10
71	1,223	556	45.46	1,660	634	38.19
77	1,868	876	46.90	3,712	1,171	31.55
94	2,000	1,955	97.75	5,953	4,368	73.37
99	1,536	456	26.69	2,527	520	20.58

Table 1. Accuracy of GPT-4o Decompositions Across Selected SMART-101 Puzzles. For each puzzle, Q_F and B_F denote the total number of sub-questions and branches, respectively, for which GPT-4o generated a final answer. Q_C and B_C denote the number of correct answers. Q_{acc} and B_{acc} represent the corresponding accuracy rates at the sub-question and branch levels.

Fine-tuning details. We fine-tuned two pretrained MLLMs: LLaVA-v1.6-Mistral-7B [29, 30] and InternVL2-8B [7, 8]. Training was performed on 4×A100 GPUs (80GB each) for 5 epochs using a learning rate of 1e-4. We adopted LoRA with a rank of 128 and an alpha value of 256. Fine-tuning was applied at each stage to evaluate the impact on model performance.

Baseline models and approaches. To benchmark the effectiveness of our proposed approach, we evaluate against four representative baseline approaches applied to both

Puzzle ID	GPT-4o	LLaVA-v1.6-Mistral-7B)				InternVL2-8B				
		Pre-DP	Pre-CoT	FT-DP	FT-CoT	Pre-DP	Pre-CoT	FT-DP	FT-CoT	ICoV
Puzzle 18	21.81	21.33	21.33	21.67	21.33	20.33	21.00	21.67	17.00	**52.67**
Puzzle 69	20.10	20.00	16.33	21.00	21.00	19.00	19.67	23.00	22.33	**26.00**
Puzzle 99	20.58	21.00	18.00	16.33	18.00	20.33	23.67	**36.67**	35.33	30.67

Table 2. Test accuracy across SMART puzzles. GPT-4o is evaluated in inference-only, direct prompting mode. "Pre" and "FT" refer to models in their pre-trained and fine-tuned states, respectively. "DP" (Direct prompting) indicates that models are prompted to directly select the final answer, while "CoT" (Chain-of-Thought) denotes that models are prompted to generate intermediate reasoning before answer selection.

LLaVA-v1.6-Mistral-7B [29, 30] and InternVL2-8B [7, 8] models. In addition, we include GPT-4o [25] as a state-of-the-art reference baseline. The four baseline variants are described as follows:

(i) *Pre-trained direct prompting (DP)*: This standard setting evaluates the model by directly presenting it with the SMART-101 puzzle question, without any guidance or intermediate reasoning steps. The model's performance reflects its zero-shot capabilities based purely on pre-trained knowledge, and is measured by its accuracy in selecting the correct option (O_{acc}). (ii) *Pre-trained explicit CoT*: Following Kojima et al. [26], this setting adopts a conventional CoT prompting strategy, encouraging the model to generate step-by-step rationales before arriving at an answer. No additional decomposition or fine-tuning is introduced in this approach. (iii) *Finetuned direct prompting (DP)*: Here, we fine-tune the model on the SMART-101 training set and evaluate it using direct prompting, i.e., without CoT prompting, similar to the pretrained DP setting. This measures the benefit of task-specific finetuning alone. (iv) *Fine–tuned explicit CoT*: Building on the finetuned direct prompting setup, this variant adds explicit CoT prompting during inference, enabling the fine-tuned model to reason through the puzzle in a structured, step-by-step manner.

These baselines provide a comprehensive evaluation of both pre-trained and fine-tuned capabilities, with and without reasoning scaffolds, across multiple state-of-the-art multimodal models.

Evaluation setup. We follow the evaluation setup from the original SMART-101 paper [15], adopting the *Instance Split* strategy. Specifically, for each root puzzle, we partition its instances into 80% for training, 5% for validation, and 15% for testing. Model performance is assessed on the test set using option selection accuracy (O_{acc}), which measures how often the model selects the correct answer from five provided options.

Our experiments focus on Puzzle 18, 69, and 99, three puzzles that exhibit the lowest branch accuracy under GPT-4o's decomposition. This indicates their higher reasoning difficulty, making them well-suited for evaluating the effectiveness of ICoV.

4.2. Empirical Results

Table 2 summarizes the test accuracy across three SMART puzzles, including Puzzle 18 (ordering), Puzzle 69 (spatial reasoning), and Puzzle 99 (counting), for GPT-4o [25], LLaVA-v1.6-Mistral-7B [29, 30], and InternVL2-8B [7, 8], under both pre-trained and fine-tuned settings, with Direct Prompting (DP) and Chain-of-Thought (CoT) prompting strategies. Across all three puzzles, standard fine-tuning on LLaVA and InternVL2 yields only marginal gains over their pre-trained counterparts. Notably, GPT-4o performs comparably. In contrast, our proposed approach ICoV, built on InternVL2-8B, demonstrates significant improvements. For Puzzle 18 (ordering), ICoV achieves a substantial boost in accuracy to 52.67%, far surpassing all baselines. Improvements are also observed on Puzzle 69 (26.00%) and Puzzle 99 (30.67%), underscoring the robustness of our approach in handling structurally complex reasoning tasks by optimizing both visual and textual reasoning.

5. Ablation Study and Discussion

5.1. Different Stages of Internalization

To assess how the internalization of reasoning stages influences performance, we conduct an ablation study on three challenging puzzles from SMART-101: Puzzle 18 (ordering), Puzzle 69 (spatial reasoning), and Puzzle 99 (counting). We fine-tune InternVL2-8B across ten progressive stages (S0 to S9), where each stage removes one additional intermediate sub-question-answer pair from training, moving from fully explicit CoT to fully implicit CoV.

The results, shown in Table 3, reveal a clear upward trend in accuracy as the model progresses from early stages toward deeper internalization. For example, Puzzle 18 achieves its highest accuracy of 52.67% at stage S8 when trained with sub-image grounding (w), significantly outperforming all baseline configurations reported in Table 2, including fine-tuned CoT-based prompting.

Interestingly, performance peaks do not always occur at the final stage. Intermediate stages such as S6–S8 of-

InternVL2-8B Option Accuracy (O_{acc}) across Fine-tuning Stages S0–S9										
Puzzle	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9
18 (w/o)	26.33	27.00	28.33	37.00	37.33	38.67	**46.33**	42.00	42.33	31.67
18 (w)	28.00	35.67	38.67	42.33	42.00	50.33	50.67	52.00	**52.67**	46.67
69 (w/o)	20.33	22.67	24.00	20.00	**26.00**	21.00	23.33	22.33	22.00	19.67
69 (w)	20.33	21.67	21.67	19.33	21.00	**24.33**	19.67	23.00	20.00	18.00
99 (w/o)	20.00	24.00	27.33	25.67	25.33	27.67	27.33	23.67	30.33	28.67
99 (w)	22.00	25.00	21.67	27.00	23.00	25.00	29.33	26.67	**30.67**	27.67

Table 3. Ablation accuracy (%) on SMART puzzles using InternVL2-8B across stage-wise fine-tuning. We evaluate the impact of ICoV's visual grounding by ablating sub-image scores used during fine-tuning. *(w/o)* indicates fine-tuning with only the generated sub-question and sub-answer (text) pairs, without access to sub-image relevance scores. *(w)* includes both sub-QA pairs and sub-image scores, enabling the model to leverage ICoV's visual grounding mechanism. Bold values highlight the best accuracy achieved per puzzle.

ten produce the highest accuracy, suggesting partial exposure to reasoning steps during training may yield stronger generalization than fully implicit reasoning alone. This trend is particularly evident in Puzzle 18, where accuracy jumps sharply between S5 (50.33%), S8 (52.67%) and S9 (46.67%).

Comparing the (w) and (w/o) variants, we find that incorporating visual grounding through sub-image attention scores consistently improves performance across all three puzzles. For instance, Puzzle 99's accuracy increases from 27.33% (S6 w/o) to 30.67% (S8 w), highlighting the effectiveness of ICoV's attention-guided fine-tuning.

5.2. Explicit CoT vs. Implicit CoV

We compare the performance of explicit Chain-of-Thought (CoT) prompting with our proposed Implicit Chain-of-Vision (ICoV) fine-tuning to understand their relative contributions to multimodal reasoning. As shown in Table 3, standard CoT prompting – whether applied to pre-trained or fine-tuned models – offers only marginal improvements over direct prompting. For instance, fine-tuned CoT on InternVL2-8B achieves 21.67%, 23.00%, and 36.67% accuracy on puzzles 18, 69, and 99, respectively—modest gains over their fine-tuned direct prompting counterparts. In contrast, ICoV yields substantial performance improvements across all puzzles, despite operating in a prompt-free, fully implicit setting. On Puzzle 18, ICOV reaches 52.67% accuracy—more than double the best CoT-based baseline. Similar gains are observed on Puzzle 69 (26.00% vs. 22.33%) and Puzzle 99 (30.67% vs. 35.33%), demonstrating ICoV's robustness across task types.

These results highlight a key advantage of ICoV: by internalizing the reasoning process during training, the model learns to autonomously perform multi-step logic without relying on external scaffolding at inference time. Unlike explicit CoT, which requires carefully engineered prompts and additional generation overhead, ICoV enables efficient and scalable inference with stronger reasoning capabilities. Overall, this comparison underscores that while explicit CoT can enhance pretrained models to some extent, ICoV fine-tuning leads to significantly more robust and generalizable reasoning—especially for structurally complex visual tasks where prompt engineering is insufficient.

6. Conclusion

In this paper, we introduce Implicit Chain-of-Vision (ICoV), a fine-tuning framework that substantially advances multimodal reasoning in large vision-language models (LVLMs). By jointly optimizing for textual and visual reasoning, ICoV enables models to internalize logical reasoning steps without relying on explicit prompting at inference time. Through a novel stepwise internalization strategy and sub-image-based attention guidance, ICoV consistently outperforms both standard fine-tuning and explicit CoT prompting methods – particularly on complex visual reasoning tasks in SMART-101 benchmark.

Our key contribution lies in bridging the gap between text-driven CoT reasoning and image-based attention mechanisms. By decomposing puzzles into structured subquestions and aligning them with localized image regions, ICoV transforms visual reasoning from a global task into a sequence of grounded and interpretable steps. This integration not only improves the answer accuracy but also enhances the model's robustness and interpretability.

Looking ahead, our goal is to make ICoV more flexible by replacing the fixed 3×3 sub-image decomposition with dynamic region selection using learnable attention. Such advancements would further extend ICoV's applicability across diverse multimodal domains beyond puzzle solving, laying the groundwork for general-purpose reasoning in real-world scenarios.

References

[1] Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*, 2024. 1

[2] Avinash Anand, Janak Kapuriya, Apoorv Singh, Jay Saraf, Naman Lal, Astha Verma, Rushali Gupta, and Rajiv Shah. Mm-phyqa: Multimodal physics question-answering with multi-image cot prompting. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 53–64. Springer, 2024. 2

[3] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17682–17690, 2024. 1

[4] Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. Large language models are visual reasoning coordinators. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[5] Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*, 2024. 1, 3

[6] Zhaorun Chen, Francesco Pinto, Minzhou Pan, and Bo Li. Safewatch: An efficient safety-policy following video guardrail model with transparent explanations. *arXiv preprint arXiv:2412.06878*, 2024. 2

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 5, 6

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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, 2024. 5, 6

[9] Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. *Advances in Neural Information Processing Systems*, 37:130185–130213, 2024. 2

[10] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024. 1, 2

[11] Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. Autoprm: Automating procedural supervision for multi-step reasoning via controllable question decomposition. *arXiv preprint arXiv:2402.11452*, 2024. 1, 3

[12] Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1254–1262, 2024. 2

[13] Zhaorun Chen, Mintong Kang, and Bo Li. Shieldagent: Shielding agents via verifiable safety policy reasoning. *arXiv preprint arXiv:2503.22738*, 2025. 2

[14] Kewei Cheng, Nesreen K Ahmed, Theodore Willke, and Yizhou Sun. Structure guided prompt: Instructing large language model in multi-step reasoning by exploring graph structure of the text. *arXiv preprint arXiv:2402.13415*, 2024. 1

[15] Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Kevin A Smith, and Joshua B Tenenbaum. Are deep neural networks smarter than second graders? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10834–10844, 2023. 1, 2, 5, 6

[16] Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*, 2024. 4

[17] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022. 1

[18] Yixiong Fang, Ziran Yang, Zhaorun Chen, Zhuokai Zhao, and Jiawei Zhou. From uncertainty to trust: Enhancing reliability in vision-language models with uncertainty-guided dropout decoding. *arXiv preprint arXiv:2412.06474*, 2024. 1

[19] Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. Question aware vision transformer for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13861–13871, 2024. 1

[20] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*, 2024. 1

[21] Buse Giledereli, Yifan Hou, Yilei Tu, and Mrinmaya Sachan. Do vision-language models really understand visual language? *arXiv preprint arXiv:2410.00193*, 2024. 1, 2

[22] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 1

[23] Ruozhen He, Paola Cascante-Bonilla, Ziyan Yang, Alexander C Berg, and Vicente Ordonez. Improved visual grounding through self-consistent explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13095–13105, 2024. 1

[24] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022. 1

[25] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6

[26] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 6

[27] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*, 2023. 1

[28] Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. Enhancing advanced visual reasoning ability of large language models. *arXiv preprint arXiv:2409.13980*, 2024. 2

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 5, 6

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 5, 6

[31] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[32] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023. 1

[33] Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18798–18806, 2024. 2

[34] Aishik Nagar, Shantanu Jaiswal, and Cheston Tan. Dissecting zero-shot visual reasoning capabilities in vision and language models. 2

[35] Aishik Nagar, Shantanu Jaiswal, and Cheston Tan. Zero-shot visual reasoning by vision-language models: Benchmarking and analysis. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024. 2

[36] Timothy Ossowski, Ming Jiang, and Junjie Hu. Prompting large vision-language models for compositional reasoning. *arXiv preprint arXiv:2401.11337*, 2024. 2

[37] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 1

[38] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*, 2024. 1

[39] Denisa Roberts and Lucas Roberts. Smart vision-language reasoners. *arXiv preprint arXiv:2407.04212*, 2024. 1

[40] Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*, 2023. 2

[41] Ayush Singh, Mansi Gupta, Shivank Garg, Abhinav Kumar, and Vansh Agrawal. Beyond captioning: Task-specific prompting for improved vlm performance in mathematical reasoning. *arXiv preprint arXiv:2410.05928*, 2024. 2

[42] Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schölkopf, and Mrinmaya Sachan. A causal framework to quantify the robustness of mathematical reasoning with language models. *arXiv preprint arXiv:2210.12023*, 2022. 2

[43] Cheng Tan, Jingxuan Wei, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Ruifeng Guo, Bihui Yu, and Stan Z Li. Boosting the power of small multimodal reasoning models to match larger models with self-consistency training. *arXiv preprint arXiv:2311.14109*, 2023. 2

[44] Chaojie Wang, Yanchen Deng, Zhiyi Lv, Shuicheng Yan, and An Bo. Q*: Improving multi-step reasoning for llms with deliberative planning. *arXiv preprint arXiv:2406.14283*, 2024. 1

[45] Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19162–19170, 2024. 2

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 1, 2

[47] Yueting Yang, Xintong Zhang, Jinan Xu, and Wenjuan Han. Empowering vision-language models for reasoning ability through large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10056–10060. IEEE, 2024. 2

[48] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 1

[49] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[50] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 1

[51] Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. *arXiv preprint arXiv:2305.14985*, 2023. 1

[52] Hao Yu, Zhuokai Zhao, Shen Yan, Lukasz Korycki, Jianyu Wang, Baosheng He, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, and Hanchao Yu. Cafe: Unifying representation and gen-

eration with contrastive-autoregressive finetuning. *arXiv preprint arXiv:2503.19900*, 2025. 1

[53] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023. 1

[54] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[55] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024. 2

[56] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*, 2024. 1

[57] Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zenghui Ding, Xianjun Yang, and Yining Sun. Beyond training: Dynamic token merging for zero-shot video understanding. *arXiv preprint arXiv:2411.14401*, 2024. 1

[58] Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zhili Feng, Zenghui Ding, and Yining Sun. Rankclip: Ranking-consistent language-image pretraining. *arXiv preprint arXiv:2404.09387*, 2024. 1

[59] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 1, 2

[60] Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[61] Zhuokai Zhao, Harish Palani, Tianyi Liu, Lena Evans, and Ruth Toner. Multimodal guidance network for missing-modality inference in content moderation. In *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–4. IEEE, 2024. 1