Let's Think with Images Efficiently! An Interleaved-Modal Chain-of-Thought Reasoning Framework with Dynamic and Precise Visual Thoughts

Xu Liu[1,2,3*], **Yongheng Zhang**[2*], **Qiguang Chen**[2], **Yao Li**[4], **Sheng Wang**[4], **Libo Qin**[1,2,3]

[1]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen
[2]School of Computer Science and Engineering, Central South University
[3]Text Computing and Cognitive Intelligence Ministry of Education Engineering Research Center, Guizhou University
[4]Shanghai Aviation Electric Co., Ltd, Aviation Industry Corporation of China, Shanghai

Abstract

Recently, Interleaved-modal Chain-of-Thought (ICoT) reasoning has achieved remarkable success by leveraging both multimodal inputs and outputs, attracting increasing attention. While achieving promising performance, current ICoT methods still suffer from two major limitations: (1) *Static Visual Thought Positioning*, which statically inserts visual information at fixed steps, resulting in inefficient and inflexible reasoning; and (2) *Broken Visual Thought Representation*, which involves discontinuous and semantically incoherent visual tokens. To address these limitations, we introduce Interleaved-modal Chain-of-Thought reasoning with **D**ynamic **a**nd **P**recise Visual Thoughts (DAP-ICoT), which incorporates two key components: (1) *Dynamic Visual Thought Integration* adaptively introduces visual inputs based on reasoning needs, reducing redundancy and improving efficiency. (2) *Precise Visual Thought Guidance* ensures visual semantically coherent and contextually aligned representations. Experiments across multiple benchmarks and models demonstrate that DAP-ICoT achieves state-of-the-art performance. In addition, DAP-ICoT significantly reduces the number of inserted images, leading to a 72.6% decrease in token consumption, enabling more efficient ICoT reasoning.

Code — https://github.com/67L1/DaP-ICoT

1 Introduction

In recent years, Multimodal Large Language Models (MLLMs) (Achiam et al. 2023; Team et al. 2024; Qin et al. 2024), and Multimodal Chain-of-Thought (MCoT) (Zhang et al. 2023; Chen et al. 2024), have significantly advanced the reasoning capabilities of MLLMs across the complex real-world tasks (Wei et al. 2025; Chen et al. 2025a). However, current MCoT mainly follows the traditional paradigm: cross-modal input, but reasoning output in the text modality, which limits the human's ability to exploit modality complementarity, reducing its reasoning performance (Fei et al. 2024; Menon, Zemel, and Vondrick 2024; Wu et al. 2025). To address this limitation, researchers propose Interleaved-Modal Chain-of-Thought (ICoT) reasoning (Hu et al. 2024; Cheng et al. 2025b; Gao et al. 2025). ICoT allows the model to receive multimodal input and simultaneously perform multimodal reasoning output, enabling visual thoughts to effectively convey the image information during reasoning (Meng et al. 2023; Shao et al. 2024; Cheng et al. 2025a; Li et al. 2025; Wang et al. 2025b; Chen et al. 2025b).

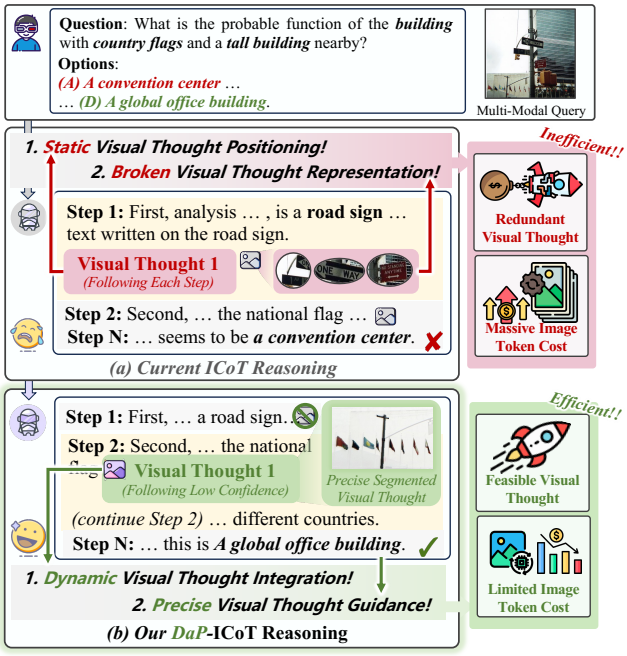

Figure 1: (a) Current ICoT: While supporting multimodal inputs and outputs, it suffers from *Static Visual Thought Integration*, which requires the insertion of visual information after each step, and *Broken Visual Thought Representation*, in which the inserted visual tokens are lack coherence, resulting in inefficient reasoning. (b) Our DAP-ICoT: It provides ***Dynamic Visual Thought Integration*** and ***Precise Visual Thought Guidance***, enabling efficient reasoning.

Specifically, a growing body of research has focused on advancing ICoT reasoning to better convey visual thoughts. Cheng et al. (2025b) employ segmentation tools to extract key areas and generate images that form visual thoughts, thereby significantly enhancing the capabilities of MLLMs for more advanced, human-like multimodal reasoning and visual operations across diverse tasks. Hu et al. (2024)

[*] Equal Contribution.

propose Sketchpad, a novel and effective sketching-based framework that enhances MLLM reasoning by enabling visual thought expression through intuitive figure drawing on a dynamic visual sketchpad interface. Zhou et al. (2024) propose Image-of-Thought visual prompting, a structured method for step-by-step visual rationale extraction that further enhances multimodal reasoning by effectively conveying internal visual thoughts in MLLMs through designed visual cues. Zhang et al. (2025a) propose Video-Text Interleaved CoT reasoning, a cognitively aligned temporal video reasoning paradigm that interleaves visual and textual information to enhance video understanding and reasoning in MLLMs. Gao et al. (2025) further employ an attention-driven selection module to statically select some image tokens that convey visual thoughts in each step.

While achieving promising performance, as shown in Figure 1 (a), current ICoT methods for visual thought conveyance face two key challenges that impede their potential:

(1) ***Static Visual Thought Positioning***: Existing ICoT reasoning approaches insert visual thoughts statically after each textual rationale generation step, resulting in rigid thinking patterns, redunant visual information, and a significant increase in computational overhead.

(2) ***Broken Visual Thought Representation***: Existing ICoT methods select discontinuous image tokens as visual thoughts, which are broken and lack coherence. Such broken visual thought impairs understanding and increases the risk of overlooking critical information.

Motivated by these challenges, we introduce an Interleaved-modal Chain-of-Thought reasoning framework with **D**ynamic **a**nd **P**recise Visual Thoughts (DAP-ICOT). Specifically, to address the first challenge, as illustrated in Figure 1 (b), we propose ***Dynamic Visual Thought Integration***, which dynamically and adaptively leverages visual information in response to evolving reasoning needs. In contrast to statically processing all available visual data, DAP-ICOT selectively invokes visual modalities based on contextual demands, ensuring that the integration of visual cues is both timely and relevant. This dynamic mechanism enhances multimodal reasoning by reducing redundant computation and focusing on salient visual cues. To address the second challenge, we introduce ***Precise Visual Thought Guidance***, which emphasizes the integration of semantically coherent and contextually relevant visual information. Instead of relying on broken cues, it employs precise visual representations that encapsulate complete semantics, ensuring tight alignment between visual input and the reasoning trajectory. This precision-oriented design safeguards conceptual consistency and enhances interpretability and reasoning accuracy. Such two modules together minimize unnecessary visual information and better capture key visual thoughts, enabling more efficient ICoT reasoning.

Experiments conducted on multiple widely used benchmarks and MLLMs demonstrate that DAP-ICOT consistently outperforms baselines. Further in-depth analysis reveals that DAP-ICOT significantly reduces the number of image insertions, resulting in a 72.6% reduction in token consumption compared to the current ICoT approach. These results highlight its effective reasoning, balancing efficiency and effectiveness in multimodal understanding.

Our contributions can be summarized as follows:

(1) We highlight two drawbacks in the existing ICoT paradigm: *Static Visual Thought Positioning* that induces inefficient reasoning, and *Broken Visual Thought Representation* that undermines coherent visual thoughts.

(2) We introduce the Interleaved-modal Chain-of-Thought reasoning with **D**ynamic **a**nd **P**recise Visual Thoughts (DAP-ICOT) to address drawbacks in ICoT, which incorporates modules: *Dynamic Visual Thought Integration* and *Precise Visual Thought Guidance*.

(3) Extensive experiments demonstrate that DAP-ICOT achieves the state-of-the-art performance. In addition, DAP-ICOT effectively reduces the number of inserted images and achieves a 72.6% reduction in token consumption, enabling more efficient reasoning.

2 DAP-ICOT Reasoning

In this work, we introduce the Interleaved-modal Chain-of-Thought reasoning with **D**ynamic **a**nd **P**recise Visual Thoughts (DAP-ICOT) to address the inefficiencies in previous ICoT approaches. Specifically, as shown in Figure 2, DAP-ICOT comprises ***Dynamic Visual Thought Integration*** (§2.1) and ***Precise Visual Thought Guidance*** (§2.2).

2.1 Dynamic Visual Thought Integration

To fill the gap of the static visual thought positioning in existing ICoT reasoning methods, we introduce ***Dynamic Visual Thought Integration*** (DVTI), as shown in Figure 2 (a), a confidence-aware strategy that dynamically decides whether visual thought should be integrated during reasoning, based on the confidence of MLLMs.

Confidence Estimation via Logit Margin Analysis

Formally, given a generated textual rationale T_t at reasoning step t, we estimate the model's confidence by analyzing token-level logit differentials during generation. This allows us to internally assess certainty without requiring external calibration. At each decoding position i, we define the local confidence δ_i as the difference between the highest and the second-highest predicted logits:

$$\delta_i = \ell_{i,w^{(1)}} - \ell_{i,w^{(2)}}, \quad \forall i \in 1, \ldots, |T_t|, \tag{1}$$

where $\ell_{i,w^{(1)}}$ and $\ell_{i,w^{(2)}}$ denote the top-1 and top-2 logits predicted at position i, respectively. This margin reflects the model's decisiveness at each token position, larger values indicate stronger confidence in the token selection.

To obtain a more robust and reliable overall confidence score for the entire rationale T_t, we aggregate the local margins by computing their mean value across positions:

$$C_t = \frac{1}{|T_t|} \sum_{i=1}^{|T_t|} \delta_i = \frac{1}{|T_t|} \sum_{i=1}^{|T_t|} \left(\ell_{i,w^{(1)}} - \ell_{i,w^{(2)}}\right), \tag{2}$$

where C_t represents the average confidence across decoding steps, with each δ_i quantifying the model's certainty at position i based on the logit gap between its top two predicted tokens, reflecting how decisively the selection of MLLMs.

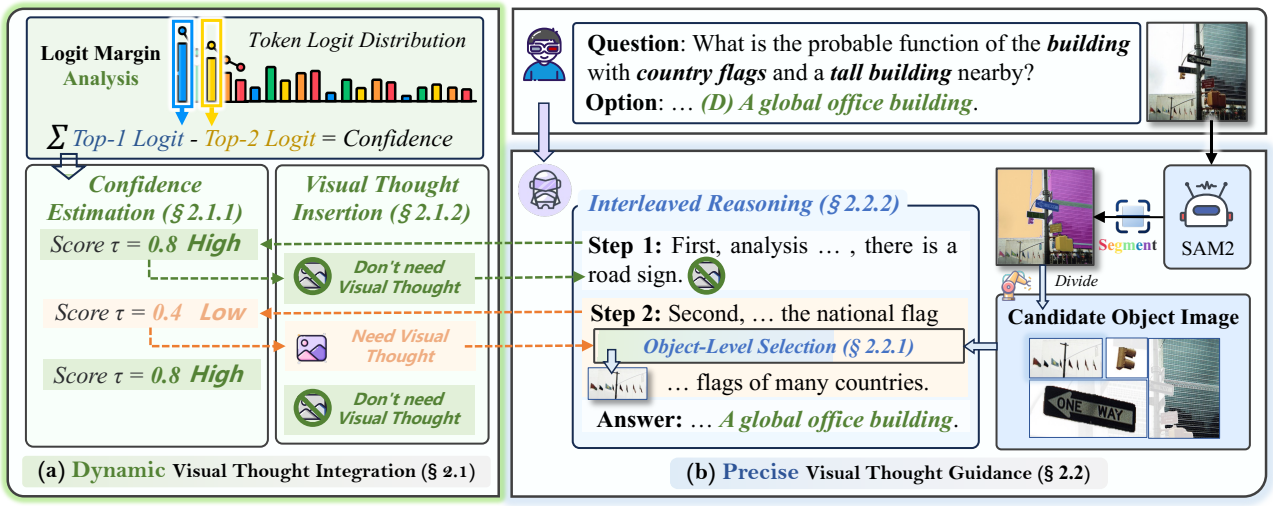

Figure 2: An overview of Interleaved-modal Chain-of-Thought reasoning with **D**ynamic **a**nd **P**recise Visual Thoughts (DAP-ICoT), including *Dynamic Visual Thought Integration* (§2.1), and *Precise Visual Thought Guidance* (§2.2).

Visual Thought Insertion Guided by Confidence

Based on the computed confidence score C_t, we apply a thresholding mechanism to determine whether visual input is necessary for the next step. The determinant is as follows:

$$I_{t+1} = \begin{cases} I^{\text{vision}}, & \text{if } C_t < \tau \\ \emptyset, & \text{otherwise} \end{cases} \tag{3}$$

where I_{t+1} denotes the visual thought input used at the $(t+1)$-th reasoning step. If the confidence C_t falls below the predefined threshold τ, the model is prompted to incorporate visual context, retrieved via the *Precise Visual Thought Guidance* (§2.2). Otherwise, no visual input is provided ($\emptyset$), allowing for proceeding based on textual reasoning.

DVTI is capable of reducing redundant information through dynamic visual information integration, thereby decreasing token processing and improving efficiency.

2.2 Precise Visual Thought Guidance

To fill the gap of broken and semantically incoherent visual inputs in existing ICoT methods, we propose ***Precise Visual Thought Guidance*** (PVTG), as shown in Figure 2 (b), which facilitates fine-grained object-level visual selection through cross-modal semantic relevance matching.

Object-Level Selection via Cross-Modal Relevance

In the first, we apply the Segment Anything Model 2 (SAM2) (Ravi et al. 2024) to the original image I_{ori} for object-level segmentation. This identifies multiple distinct object regions, each corresponding to semantically meaningful visual content. We then extract sub-images of objects, forming a pool of candidate object-centric visual inputs. Unlike patch-level image tokens, these object-centric sub-images preserve semantic information. Specifically:

$$\mathcal{O} = O_1, O_2, \ldots, O_N, \tag{4}$$

where each O_i denotes a complete and semantically coherent object sub-image extracted from the original image I_{ori}.

When *Dynamic Visual Thought Integration* (§2.1) identifies the need for visual input at step t, we compute the semantic relevance between the textual rationale T_t and each candidate object image $O_i \in \mathcal{O}$ via cross-modal attention using a similarity function $f_{\text{attn}}(\cdot,\cdot)$:

$$s_i = f_{\text{attn}}(T_t, O_i), \tag{5}$$

where s_i denotes the attention-based similarity between rationale T_t and object image O_i (Gao et al. 2025). We then select the most relevant object image with the highest score:

$$\hat{O} = \underset{O_i \in \mathcal{O}}{\operatorname{argmax}}\, s_i. \tag{6}$$

where $\hat{O}$ denotes the object image that exhibits the strongest semantic alignment with the current textual rationale.

Interleaved Reasoning with Aligned Visual Inputs

Instead of treating $\hat{O}$ as a standalone input, we interleave it into the reasoning process by embedding it within the textual rationale $\mathcal{R}_t$, forming a multimodal reasoning sequence:

$$\mathcal{R}_{t\to v} = \mathcal{R}_t \oplus \hat{O}, \tag{7}$$

where $\oplus$ denotes the operation of embedding the selected object image $\hat{O}$ into the textual reasoning sequence $\mathcal{R}_t$, specifically by inserting the image token after the previously generated textual rationale.

The model then continues the next reasoning step based on this interleaved multimodal input:

$$\mathcal{R}_{t+1} = \underset{\mathcal{R}}{\operatorname{argmax}}\, P(\mathcal{R} \mid Q,\ \mathcal{R}_{t\to v}, P_{t\to v}), \tag{8}$$

where Q denotes the question, and $P_{t\to v}$ represents the prompt constructed for interleaved reasoning.

Through targeted selection and structured interleaving, PVTG provides precise visual information that preserves semantic coherence, reduces noise, and enhances the effectiveness and interpretability of the multimodal reasoning.

Models	Methods	M^{3}CoT		ScienceQA		MME	
		0-Shot Acc. ↑	1-Shot Acc. ↑	0-Shot Acc. ↑	1-Shot Acc. ↑	0-Shot Score ↑	1-Shot Score ↑
Chameleon-7B (Team 2024)	DIRECT	22.5	23.8	43.1	43.4	724.2	942.9
	MMCOT TMLR 2024	26.0	28.0	46.2	48.9	435.8	661.2
	DDCOT NeurIPS 2023	29.8	30.3	47.4	48.4	725.9	953.8
	SCAFFOLD ACL 2025	31.0	31.2	48.6	50.7	388.1	634.5
	CCOT CVPR 2024	25.1	26.3	42.8	44.3	366.1	487.9
	ICOT CVPR 2025	26.1	32.1	44.5	45.3	794.8	928.9
	DAP-ICOT	**41.0**	**41.9**	**57.1**	**62.9**	**832.3**	**1013.0**
LLaVA-V1.5-7B (Liu et al. 2023)	DIRECT	23.2	25.7	21.9	23.5	975.3	1079.8
	MMCOT TMLR 2024	29.6	30.4	42.4	43.1	1140.2	1277.4
	DDCOT NeurIPS 2023	25.4	26.5	29.3	31.0	736.7	991.8
	SCAFFOLD ACL 2025	26.6	28.3	37.7	39.8	1170.3	1264.3
	CCOT CVPR 2024	34.2	35.5	33.8	35.7	1294.3	1349.5
	ICOT CVPR 2025	34.6	35.0	41.7	46.7	1331.6	1421.6
	DAP-ICOT	**36.3**	**37.6**	**50.4**	**51.1**	**1386.7**	**1526.9**
LLaVA-V1.5-13B (Liu et al. 2023)	DIRECT	24.6	25.5	29.7	34.0	995.4	1118.5
	MMCOT TMLR 2024	32.1	32.9	56.1	58.3	1078.1	1224.3
	DDCOT NeurIPS 2023	32.9	33.8	39.3	41.8	800.9	1034.1
	SCAFFOLD ACL 2025	31.9	33.2	41.7	43.8	1231.2	1389.9
	CCOT CVPR 2024	30.1	32.0	45.0	46.3	1249.3	1442.3
	ICOT CVPR 2025	37.0	37.9	54.6	54.8	1405.4	1523.8
	DAP-ICOT	**39.4**	**41.8**	**60.3**	**62.7**	**1556.3**	**1726.3**
Qwen2-VL-2B (Wang et al. 2024)	DIRECT	14.4	24.5	64.3	65.2	641.5	741.3
	MMCOT TMLR 2024	14.9	22.4	65.6	67.3	1102.8	1304.7
	DDCOT NeurIPS 2023	37.9	39.3	65.8	68.4	800.6	967.0
	SCAFFOLD ACL 2025	40.3	43.6	66.7	69.4	1344.8	1536.2
	CCOT CVPR 2024	20.2	37.7	64.2	66.5	761.6	867.5
	ICOT CVPR 2025	35.8	37.3	60.4	67.0	941.9	1453.9
	DAP-ICOT	**47.3**	**51.0**	**68.4**	**73.6**	**1378.9**	**1862.4**
Qwen2-VL-7B (Wang et al. 2024)	DIRECT	33.0	35.0	70.9	71.2	1599.3	1641.3
	MMCOT TMLR 2024	44.4	47.5	70.8	73.8	1602.5	1874.3
	DDCOT NeurIPS 2023	43.9	45.3	62.8	65.3	1752.4	1826.6
	SCAFFOLD ACL 2025	49.9	53.6	74.4	75.0	1668.2	1822.2
	CCOT CVPR 2024	48.7	53.0	72.7	74.8	1866.3	1941.9
	ICOT CVPR 2025	38.0	44.8	54.2	67.0	1587.3	1709.3
	DAP-ICOT	**57.2**	**58.7**	**75.9**	**78.5**	**2012.2**	**2076.0**

Table 1: The main experimental results. **Bold** indicates the best performance. For the M^{3}CoT and ScienceQA, Acc. is used as the evaluation metric, while for the MME, the sum of the Perception and Cognition scores is used as the evaluation metric.

3 Experiments and Analysis

3.1 Experiments Setting

Following Gao et al. (2025), in addition to the direct query approach, we also adopt the following methods as baselines:

- `MMCoT` (Zhang et al. 2024) separates rationale generation and answer inference by incorporating both text and image modalities to improve reasoning performance.
- `DDCoT` (Zheng et al. 2023) divides reasoning and recognition by combining LLM reasoning with visual recognition through negative-space prompting, enabling effective and explainable multimodal CoT reasoning.
- `SCAFFOLD` (Lei et al. 2025) prompts overlays a dot matrix on images as visual anchors and introduces coordinate-based textual references to effectively enhance vision-language coordination in MLLMs.
- `CCoT` (Mitra et al. 2024) first generates scene graphs with LMMs and then uses them in prompts to enhance compositional reasoning without annotations.
- `ICoT` (Gao et al. 2025) generates paired visual and textual reasoning steps by inserting image regions via Attention-driven Selection to enhance reasoning.

All methods are reproduced once using their official open-source implementations and evaluated under both 0-shot and 1-shot settings. To evaluate the effectiveness of DAP-ICOT, we conduct experiments on five MLLMs, including Chameleon-7B (Team 2024), LLaVA-V1.5-(7B, 13B) (Liu et al. 2023), and Qwen2-VL-(2B, 7B) (Wang et al. 2024). We use the default top-p and temperature settings provided by each MLLM. In DAP-ICOT, the confidence threshold τ is tuned on M^{3}CoT validation set by searching within [0, 1] and selecting the value that yields the best performance.

3.2 Main Results

The experimental results are summarized in Table 1. Based on these results, we can observe the following:

(1) **DAP-ICOT achieves consistently superior performance.** DAP-ICOT consistently achieves the highest reasoning accuracy across all settings. In particular, it significantly outperforms all baseline methods on the M^3CoT benchmarks under both 0-shot and 1-shot settings. For example, on M^3CoT task with the Chameleon-7B, DAP-ICOT achieves a remarkable 0-shot accuracy of 41.0%, which is substantially higher than the second-best method, Scaffold, with a score of 31.0%.

(2) **DAP-ICOT demonstrates strong versatility across diverse tasks.** In addition to its promising performance on M^3CoT, DAP-ICOT consistently outperforms all baselines across other challenging reasoning tasks and comprehensive multimodal benchmarks. Specifically, it achieves the highest average scores on ScienceQA and MME in all settings. This demonstrates its strong generalization ability and adaptability to complex reasoning and holistic multimodal understanding tasks.

(3) **DAP-ICOT is generalizable across MLLMs of different architectures and scales.** DAP-ICOT consistently delivers superior performance across various MLLMs with diverse architectures and scales. From small MLLMs such as *Qwen2-VL-2B* to large MLLMs like *Qwen2-VL-7B* and *LLaVA-V1.5-13B*, DAP-ICOT maintains its leading performance. These results clearly suggest that DAP-ICOT is not only effective for specific models but also adaptable to a wide range of pretraining settings and reasoning capacities, demonstrating scalability and model-agnostic robustness.

3.3 Analysis

This section provides a more in-depth analysis of DAP-ICOT to demonstrate its effectiveness and efficiency.

Both the DVTI and PVTG modules are vital for addressing key ICoT challenges. We perform thorough ablation studies on *Qwen2-VL-7B* to systematically assess the impact of *Dynamic Visual Thought Integration* (DVTI) and *Precise Visual Thought Guidance* (PVTG). The results, as shown in Figure 3, clearly confirm their effectiveness.

- Removing the *Dynamic Visual Thought Integration* (DVTI) module results in a substantial performance degradation—specifically, a 14.4% drop on the M^3CoT benchmark and a 20.8% drop on the ScienceQA benchmark. These results underscore the critical role of DVTI in dynamically fusing multimodal information.
- Removing the *Precise Visual Thought Guidance* (PVTG) module leads to a significant performance decline, with a 13.8% reduction on the M^3CoT benchmark and 20.4% on the ScienceQA benchmark. This highlights the importance of PVTG in structuring the visual input space by leveraging fine-grained object-level segmentation.

These findings demonstrate that both DVTI and PVTG are essential for efficient and precise ICoT reasoning.

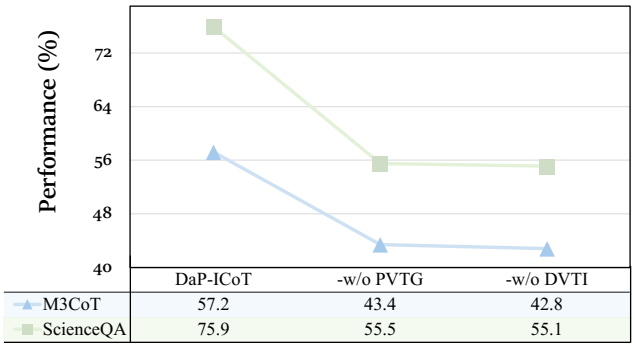

Figure 3: Ablation Study on *Qwen2-VL-7B*: "w/o PVTG" indicates removal of *Precise Visual Thought Guidance* for Visual Cues, and "w/o DVTI" indicates removal of *Dynamic Visual Thought Integration* for Adaptive Reasoning

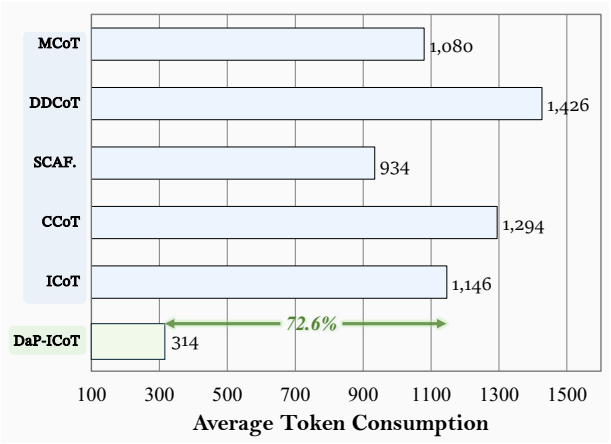

Figure 4: A comparison of the total token consumption between DAP-ICOT and baseline methods on the M^3CoT benchmark using the *Qwen2-VL-7B*. DAP-ICOT achieves a 72.6% reduction in token consumption compared to ICoT.

DAP-ICOT significantly reduces the token consumption of MLLMs. To further verify the lightweight design of DAP-ICOT, we compare its total token consumption with several baseline methods on the M^3CoT using the *Qwen2-VL-7B*. This experiment aims to assess whether DAP-ICOT can effectively reduce token usage while maintaining strong reasoning performance. As shown in Figure 4, DAP-ICOT achieves a significant reduction in token consumption, using an average of 314 tokens, which is a 72.6% decrease compared to `ICoT` that requires 1,146 tokens. Furthermore, most baselines consume considerably more tokens, with `CCoT` requiring 1,294 tokens and `DDCoT` reaching the highest at 1,426 tokens. Even relatively more efficient methods, such as `MMCoT` and `SCAFFOLD`, still require 1,080 and 934 tokens, respectively. These results indicate that the *Dynamic Visual Thought Integration* module in DAP-ICOT effectively reduces the overhead associated with image embedding, thereby achieving significantly lower token consumption compared to existing baseline reasoning methods.

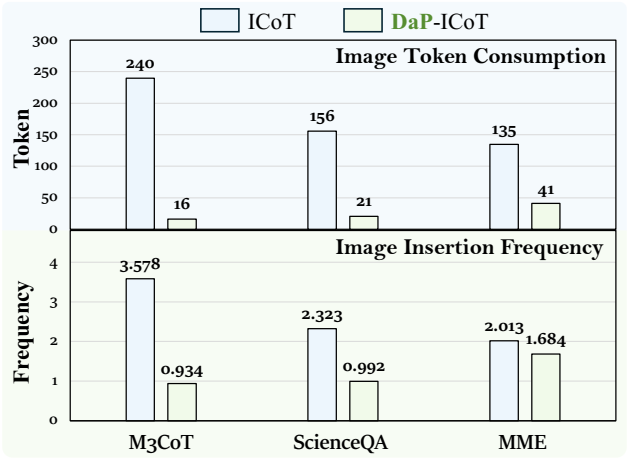

Figure 5: A comparison of image insertion frequency and the number of inserted image tokens between DAP-ICOT and ICoT on the *Qwen2-VL-7B* model.

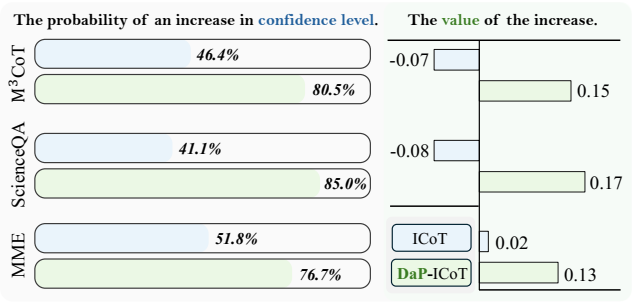

Figure 6: The proportion of samples where the confidence level increases after image insertion, as well as the average confidence value improvement, for DAP-ICOT and ICoT. The Blue Color denotes the baseline method ICoT, while the Green Color represents our DAP-ICOT.

DAP-ICOT reduces resource consumption from image insertions. To clarify the efficient feature of DAP-ICOT, we conduct a comparative analysis with ICoT on image usage during reasoning, based on two key metrics: (1) The average number of image insertions per sample and (2) The average number of image tokens consumed after insertion. As shown in Figure 5, DAP-ICOT demonstrates a significant reduction in both image insertion frequency and token consumption. On average, DAP-ICOT inserts only 1.2 images per sample, whereas ICoT inserts an average of 2.6 images. Moreover, in terms of token usage, DAP-ICOT consumes merely 26 image tokens on average, which is substantially lower than ICoT. These results demonstrate the efficiency of DAP-ICOT, which achieves superior performance with minimal visual input and token consumption. This is due to its modules: *Dynamic Visual Thought Integration*, which adaptively reduces unnecessary image insertions, and *Precise Visual Thought Guidance*, which effectively minimizes token usage by selecting compact object-level visual inputs.

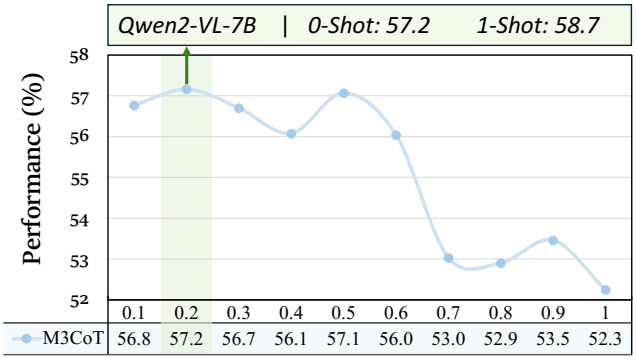

Figure 7: Performance comparison across the Confidence Threshold τ range of $(0, 1]$: Optimal model performance achieved at $\tau = 0.2$, with 57.2% Accuracy in the 0-shot setting and 58.7% in the 1-shot setting, effectively balancing visual integration and textual reasoning.

DAP-ICOT effectively enhances confidence during the reasoning process. To further investigate why DAP-ICOT achieves superior performance, we conduct an in-depth analysis of the model's reasoning confidence variations after image insertion. Specifically, we compare DAP-ICOT and ICoT in terms of two key metrics: (1) The proportion of samples showing increased confidence after applying Visual Thought, and (2) The average magnitude of the confidence improvement. As shown in Figure 6, DAP-ICOT consistently demonstrates a significantly higher probability of confidence improvement across all three benchmarks. On average, DAP-ICOT leads to an increase in confidence for 80.7% of the samples, while ICoT only improves confidence in 46.4% of cases. Furthermore, regarding the extent of confidence enhancement, DAP-ICOT consistently outperforms ICoT across all three benchmarks. These results clearly indicate that DAP-ICOT is more effective in leveraging visual information to enhance confidence during reasoning, thereby contributing to its superior performance.

The search strategy for the confidence threshold τ. To gain a clearer understanding of the threshold selection process for the confidence threshold τ in the *Dynamic Visual Thought Integration* (DVTI) module, we conduct a systematic threshold search experiment using *Qwen2-VL-7B* on the M^3CoT benchmark. Specifically, we vary the confidence threshold τ within the range of $(0, 1]$ with an interval of 0.1 and evaluate the model's performance under each setting. The experimental results are presented in Figure 7, illustrating the relationship between the confidence threshold and overall performance. Reveal the impact of varying the confidence threshold τ on model performance. As τ increases, the performance exhibits a clear trend. The optimal performance of 57.2% is achieved at a threshold of 0.2. This trend suggests that a moderate threshold effectively balances the model's reliance on visual information and textual reasoning, enabling optimal reasoning performance. In contrast, overly conservative or overly aggressive visual interactions negatively affect the model's ability to reason effectively.

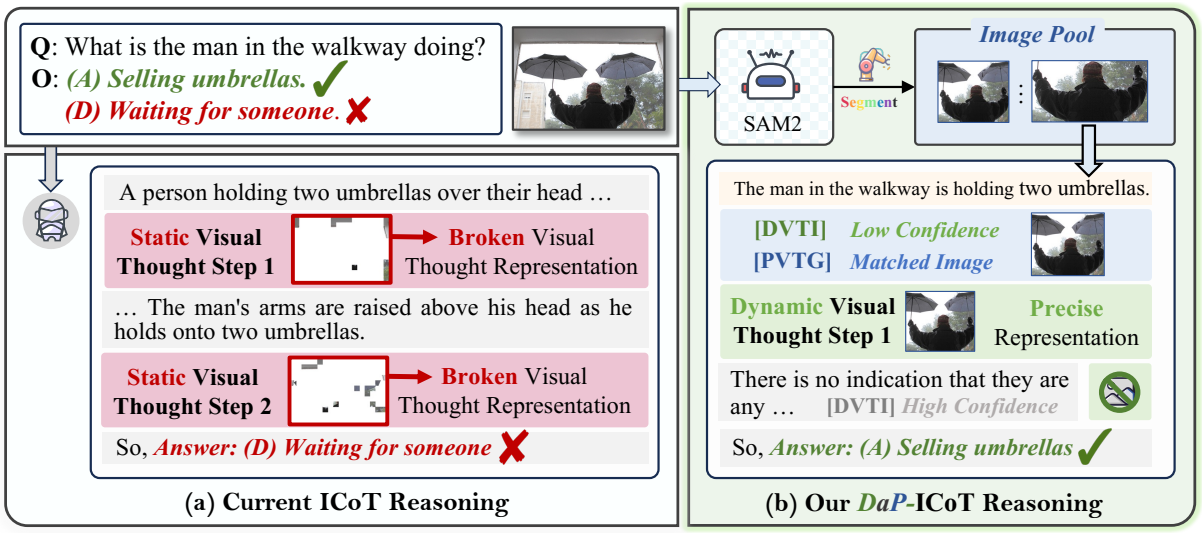

Figure 8: The case study. Figure (a) illustrates the reasoning process of the current ICoT, where after each reasoning step, it is ***necessary*** to insert image tokens that are most similar to the previous text. However, these tokens are ***broken*** image tokens, which ultimately leads to the incorrect answer *(D)*. Figure (b) illustrates our DAP-ICOT. In contrast, DAP-ICOT ***only*** inserts images when the text confidence is low. Furthermore, the inserted images are ***precise and complete***, having been segmented using the SAM2 model. After efficient interleaved visual-textual reasoning, the final correct answer *(A)* is obtained.

Qualitative Analysis. To better understand the performance of DAP-ICOT, we present a real-world example. As shown in Figure 8 (a), ICoT inserts incomplete image tokens during reasoning, leading to an incorrect result (*D*). This example highlights the limitations of the current ICoT approach and its impact on reasoning accuracy. In contrast, Figure 8 (b) demonstrates DAP-ICOT, which inserts only complete, context-relevant images when text confidence is low. Compared to ICoT, DAP-ICOT inserts *fewer* images and employs a more selective approach. After ICoT reasoning, the system correctly arrives at answer (*A*), which is verified through the dynamic visual thought integration mechanism. This example illustrates the efficiency of the selective visual input mechanism in DAP-ICOT reasoning.

4 Related Work

In recent years, Multimodal Large Language Models (MLLMs) have witnessed rapid advancements (Liang et al. 2024; Qiu et al. 2025; Qiang et al. 2025), and the emergence of Multimodal Chain-of-Thought (MCoT) reasoning has further enhanced their performance (Wang et al. 2025b). Specifically, Zhang et al. (2024) proposed Multimodal-CoT, a two-stage reasoning framework integrating text and image modalities. Chen et al. (2024) introduce M^3CoT, a benchmark for multi-modal, multi-step, and multi-domain chain-of-thought reasoning, addressing key limitations of existing MCoT benchamrk. However, existing MCoT methods largely follow the conventional paradigm of taking cross-modal inputs while generating reasoning outputs only in the text modality. This limits the effective use of modality complementarity and diminishes reasoning performance (Wang et al. 2025a; Lin et al. 2025; Zhang et al. 2025b).

To address this limitation, researchers have explored Interleaved-Modal Chain-of-Thought (ICoT) (Gao et al. 2025; Wu et al. 2025), which enhances the reasoning abilities of MLLMs through cross-modal Integrations (Hu et al. 2024; Cheng et al. 2025a). For example, Zhou et al. (2024) propose Image-of-Thought prompting to guide MLLMs in step-by-step visual extraction. Hu et al. (2024) introduce a sketching framework, enabling models to perform human-like drawing to reasoning. In addition, Cheng et al. (2025b) propose the CoMT for evaluating multimodal reasoning with visual and textual operations. Gao et al. (2025) introduce ICoT reasoning with an Attention-driven Selection for generating interleaved visual-textual reasoning.

Compared to previous ICoT reasoning approaches, DAP-ICOT introduces both *Dynamic Visual Thought Integration* and *Precise Visual Thought Guidance*, enabling not only more efficient reasoning but also adaptive and context-aware visual clues for ICoT Reasoning.

5 Conclusion

In this work, we propose Interleaved-modal Chain-of-Thought reasoning with **D**ynamic **a**nd **P**recise Visual Thoughts (DAP-ICOT), achieving efficient reasoning. Specifically, DAP-ICOT adaptively integrates informative and context-relevant visual information and provides semantically coherent visual inputs. Extensive evaluations on multiple benchmarks and advanced MLLMs demonstrate that DAP-ICOT achieves superior performance. In addition, DAP-ICOT is capable of effectively reducing token consumption and the frequency of visual insertions, highlighting its strong potential in efficient multimodal reasoning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) via grants 92570120 and 62306342. This work was supported by the Scientific Research Fund of Hunan Provincial Education Department (24B0001). This work was sponsored by the Excellent Young Scientists Fund in Hunan Province (2024JJ4070), the Science and Technology Innovation Program of Hunan Province under Grant 2024RC3024, and CCF-Zhipu Large Model Innovation Fund (NO.CCF-Zhipu202406). This study was also funded by the Open Project of the Text Computing and Cognitive Intelligence Ministry of Education Engineering Research Center (No. TCCI250101). Libo Qin is the corresponding author.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; ; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chen, Q.; Qin, L.; Liu, J.; Peng, D.; Guan, J.; Wang, P.; Hu, M.; Zhou, Y.; Gao, T.; and Che, W. 2025a. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Chen, Q.; Qin, L.; Zhang, J.; Chen, Z.; Xu, X.; and Che, W. 2024. M^3CoT: A Novel Benchmark for Multi-Domain Multi-step Multi-modal Chain-of-Thought. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8199–8221. Bangkok, Thailand: Association for Computational Linguistics.

Chen, Q.; Yang, M.; Qin, L.; Liu, J.; Yan, Z.; Guan, J.; Peng, D.; Ji, Y.; Li, H.; Hu, M.; et al. 2025b. AI4Research: A Survey of Artificial Intelligence for Scientific Research. *arXiv preprint arXiv:2507.01903*.

Cheng, Z.; Chen, Q.; Xu, X.; Wang, J.; Wang, W.; Fei, H.; Wang, Y.; Wang, A. J.; Chen, Z.; Che, W.; et al. 2025a. Visual Thoughts: A Unified Perspective of Understanding Multimodal Chain-of-Thought. *arXiv preprint arXiv:2505.15510*.

Cheng, Z.; Chen, Q.; Zhang, J.; Fei, H.; Feng, X.; Che, W.; Li, M.; and Qin, L. 2025b. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23678–23686.

Fei, H.; Wu, S.; Ji, W.; Zhang, H.; Zhang, M.; Lee, M.-L.; and Hsu, W. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*.

Gao, J.; Li, Y.; Cao, Z.; and Li, W. 2025. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19520–19529.

Hu, Y.; Shi, W.; Fu, X.; Roth, D.; Ostendorf, M.; Zettlemoyer, L.; Smith, N. A.; and Krishna, R. 2024. Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Lei, X.; Yang, Z.; Chen, X.; Li, P.; and Liu, Y. 2025. Scaffolding Coordinates to Promote Vision-Language Coordination in Large Multi-Modal Models. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 2886–2903. Abu Dhabi, UAE: Association for Computational Linguistics.

Li, C.; Wu, W.; Zhang, H.; Xia, Y.; Mao, S.; Dong, L.; Vulić, I.; and Wei, F. 2025. Imagine while Reasoning in Space: Multimodal Visualization-of-Thought. *arXiv preprint arXiv:2501.07542*.

Liang, Z.; Xu, Y.; Hong, Y.; Shang, P.; Wang, Q.; Fu, Q.; and Liu, K. 2024. A Survey of Multimodel Large Language Models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, 405–409.

Lin, J.; Zeng, X.; Zhu, J.; Wang, S.; Shun, J.; Wu, J.; and Zhou, D. 2025. Plan and Budget: Effective and Efficient Test-Time Scaling on Large Language Model Reasoning. *arXiv preprint arXiv:2505.16122*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Meng, F.; Yang, H.; Wang, Y.; and Zhang, M. 2023. Chain of images for intuitively reasoning. *arXiv preprint arXiv:2311.09241*.

Menon, S.; Zemel, R.; and Vondrick, C. 2024. Whiteboard-of-Thought: Thinking Step-by-Step Across Modalities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 20016–20031.

Mitra, C.; Huang, B.; Darrell, T.; and Herzig, R. 2024. Compositional Chain-of-Thought Prompting for Large Multimodal Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14420–14431. IEEE Computer Society.

Qiang, C.; Wei, Z.; Han, X.; Wang, Z.; Li, S.; Lan, X.; Jiao, J.; and Han, Z. 2025. VER-Bench: Evaluating MLLMs on Reasoning with Fine-Grained Visual Evidence. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 12698–12705.

Qin, L.; Chen, Q.; Feng, X.; Wu, Y.; Zhang, Y.; Li, Y.; Li, M.; Che, W.; and Yu, P. S. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

Qiu, T.; Gao, J.; Li, J.; Leong, H.; Huang, X.; Wang, X.; Zhang, X.; Xu, K.; and Zhang, L. 2025. Intentvcnet: Bridging spatio-temporal gaps for intention-oriented controllable video captioning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 13822–13829.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. arXiv:2408.00714.

Shao, H.; Qian, S.; Xiao, H.; Song, G.; Zong, Z.; Wang, L.; Liu, Y.; and Li, H. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37: 8612–8642.

Team, C. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; Mariooryad, S.; Ding, Y.; Geng, X.; Alcober, F.; Frostig, R.; Omernick, M.; Walker, L.; Paduraru, C.; Sorokin, C.; Tacchetti, A.; Gaffney, C.; Daruki, S.; Sercinoglu, O.; Gleicher, Z.; Love, J.; Voigtlaender, P.; Jain, R.; Surita, G.; Mohamed, K.; Blevins, R.; Ahn, J.; Zhu, T.; Kawintiranon, K.; Firat, O.; Gu, Y.; Zhang, Y.; Rahtz, M.; Faruqui, M.; Clay, N.; Gilmer, J.; Co-Reyes, J.; Penchev, I.; Zhu, R.; Morioka, N.; Hui, K.; Haridasan, K.; Campos, V.; Mahdieh, M.; Guo, M.; Hassan, S.; Kilgour, K.; Vezer, A.; Cheng, H.-T.; de Liedekerke, R.; Goyal, S.; Barham, P.; Strouse, D.; Noury, S.; Adler, J.; Sundararajan, M.; Vikram, S.; Lepikhin, D.; Paganini, M.; Garcia, X.; Yang, F.; Valter, D.; Trebacz, M.; Vodrahalli, K.; Asawaroengchai, C.; Ring, R.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Wang, B.; Li, Y.; Zhou, Q.; Leong, H. Y.; Zhao, T.; Ye, L.; Deng, H.; Luo, D.; and Vasconcelos, N. 2025a. Do Vision Language Models infer human intention without visual perspective-taking? Towards a scalable" One-Image-Probe-All" dataset.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; Lin, J.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Y.; Wu, S.; Zhang, Y.; Yan, S.; Liu, Z.; Luo, J.; and Fei, H. 2025b. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.

Wei, Z.; Qiang, C.; Jiang, B.; Han, X.; Yu, X.; and Han, Z. 2025. AD^ 2-Bench: A Hierarchical CoT Benchmark for MLLM in Autonomous Driving under Adverse Conditions. *arXiv preprint arXiv:2506.09557*.

Wu, X.; Liu, J.; Huang, D.; Li, X.; Wang, Y.; Chen, C.; Ma, L.; Cao, X.; and Xue, J. 2025. ViC-Bench: Benchmarking Visual-Interleaved Chain-of-Thought Capability in MLLMs with Free-Style Intermediate State Representations. *arXiv preprint arXiv:2505.14404*.

Zhang, Y.; Liu, X.; Tao, R.; Chen, Q.; Fei, H.; Che, W.; and Qin, L. 2025a. ViTCoT: Video-Text Interleaved Chain-of-Thought for Boosting Video Understanding in Large Language Models. *arXiv preprint arXiv:2507.09876*.

Zhang, Y.; Liu, X.; Zhou, R.; Chen, Q.; Fei, H.; Lu, W.; and Qin, L. 2025b. CCHall: A Novel Benchmark for Joint Cross-Lingual and Cross-Modal Hallucinations Detection in Large Language Models. *arXiv preprint arXiv:2505.19108*.

Zhang, Z.; Zhang, A.; Li, M.; Karypis, G.; Smola, A.; et al. 2024. Multimodal Chain-of-Thought Reasoning in Language Models. *Transactions on Machine Learning Research*.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zheng, G.; Yang, B.; Tang, J.; Zhou, H.-Y.; and Yang, S. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36: 5168–5191.

Zhou, Q.; Zhou, R.; Hu, Z.; Lu, P.; Gao, S.; and Zhang, Y. 2024. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*.