CMMCoT: Enhancing Complex Multi-Image Comprehension via Multi-Modal Chain-of-Thought and Memory Augmentation

Guanghao Zhang[1*], Tao Zhong[1*], Yan Xia[1,2*], Mushui Liu[1,2*], Zhelun Yu[1], Haoyuan Li[1], Wanggui He[1], Dong She[1], Yi Wang[1,2], Hao Jiang[1†]

[1]Alibaba Group, China
[2] College of Computer Science and Technology, Zhejiang University, China
{guanghao.zgh, zt395565}@taobao.com, {xiayan.zju, lms}@zju.edu.cn, yuzhelun.yzl@taobao.com, lihaoyuan@zju.edu.cn, wanggui.hwg@taobao.com, sd0809@mail.ustc.edu.cn, y_ w@zju.edu.cn, aoshu.jh@taobao.com

Abstract

While previous multimodal slow-thinking methods have demonstrated remarkable success in single-image understanding scenarios, their effectiveness becomes fundamentally constrained when extended to more complex multi-image comprehension tasks. This limitation stems from their predominant reliance on text-based intermediate reasoning processes. While for human, when engaging in sophisticated multi-image analysis, they typically perform two complementary cognitive operations: (1) continuous cross-image visual comparison through region-of-interest matching, and (2) dynamic memorization of critical visual concepts throughout the reasoning chain. Motivated by these observations, we propose the **Complex Multi-Modal Chain-of-Thought (CMMCoT)** framework, a multi-step reasoning framework that mimics human-like "slow thinking" for multi-image understanding. Our approach incorporates two key innovations: (1) The construction of interleaved **multimodal multi-step reasoning chains**, which utilize critical visual region tokens, extracted from intermediate reasoning steps, as supervisory signals. This mechanism not only facilitates comprehensive cross-modal understanding but also enhances model interpretability. (2) The introduction of a **test-time memory augmentation** module that expands the model's reasoning capacity during inference while preserving parameter efficiency. Furthermore, to facilitate research in this direction, we have curated a novel multi-image slow-thinking dataset. Extensive experiments demonstrate the effectiveness of our model.

Introduction

Recent years have witnessed the rapid advancement of generative (He et al. 2025; Esser et al. 2024; Wang et al. 2025; Liu et al. 2025; Ma et al. 2024, 2025; Feng et al. 2025) and multimodal understanding models (Wang et al. 2025, 2024b). In particular, multi-modal large language models (MLLMs) have achieved remarkable breakthroughs across various multimodal tasks (Liu et al. 2024b; Wang et al. 2023, 2024a), such as multimodal recognition, localization, and single- as well as multi-image understanding. However, current MLLM methods employing the "direct prediction" paradigm to generate answers end-to-end exhibit two critical limitations when confronted with intricate scenarios: **(a)** They systematically overlook latent evidential features embedded within cross-modal data patterns, resulting in erroneous predictions—even in state-of-the-art models like GPT-4V(Yang et al. 2023), and **(b)** Their decision-making processes suffer from a pronounced lack of interpretability, undermining their reliability and transparency in critical applications.

The remarkable success of the O1 model has catalyzed growing research interest in chain-of-thought (CoT) reasoning and "slow thinking" mechanisms (Snell et al. 2024; Prystawski, Li, and Goodman 2024). These mechanisms yield substantial performance enhancements over conventional LLMs and significantly enhance mathematical reasoning and logical deduction capabilities. However, unimodal slow-thinking methodologies exhibit inherent constraints when directly adapted to multimodal domains, primarily due to insufficient cross-modal spatial reasoning capacities for complex scene comprehension.

Existing multimodal CoT approaches predominantly focus on employing external tools to annotate complete textual reasoning chains (Ni et al. 2024; Xu et al. 2024; Shao et al. 2024) or leveraging textual reasoning capabilities to guide visual inference (Du et al. 2025) while neglecting supervision of visual reasoning processes during training.

Human cognition in multimodal contexts, particularly when processing multi-image compositions, operates through a dual-process cognitive mechanism: (1) parallel processing of linguistic semantics and visual signal interpretation, and (2) synergistic integration of textual deductive reasoning with active visual pattern mining. This cognitive architecture enables the iterative establishment of object-relation correspondences across cross-modal representations, ultimately synthesizing a coherent cross-modal reasoning framework through dynamic interaction between linguistic parsing and visual grounding. Although previous

*These authors contributed equally.
†Corresponding authors.

works like MVoT (Li et al. 2025) and VoCoT (Li et al. 2024d) have been proposed to mimic the aforementioned process, these methods primarily focus on enhancing the slow reasoning capabilities of models in single-image scenarios, while leaving the more complex multi-image scenarios largely unexplored.

Thus in this paper, we address the challenge of slow thinking in complex multi-image scenarios. However, constructing multi-step reasoning paths for multi-image scenes presents two significant challenges: **Complexity in Cross-image Visual Concept Tracking:** Unlike single-image scenarios, comprehending multi-image scenes (Jiang et al. 2024) necessitates the correlation of visual objects across disparate images and the integration of textual information to facilitate comprehensive reasoning. **Enhancement of Model's Inference Capabilities during Testing:** While test-time scaling has shown promise in boosting model reasoning (Ni et al. 2024; Xu et al. 2024) without parameter increases, its effectiveness diminishes in complex multi-image scenarios. Furthermore, simply scaling pre-trained model parameters is reaching a performance ceiling. Thus, exploring alternative methods to enhance model capabilities during testing for multi-image understanding is crucial.

To address the aforementioned challenges, we propose a novel **Complex Multi-Modal CoT (CMMCoT)** framework that empowers models with slow-thinking capability in multi-image scenarios, significantly enhancing reasoning performance. To overcome the limitations of previous models that are constrained to generating coordinates during reasoning, we devise a novel training-inference paradigm. This paradigm enables the model to extract visual tokens of key objects, grounded by the coordinates mentioned in the reasoning chain, and subsequently predict subsequent reasoning steps and final answers.

Building upon prior works (Brown et al. 2024; Wang et al. 2024c), we further propose the **Retrieval-based Image Feature Reasoning Enhancement Module (RIFREM)** module to enable comprehensive cross-modal deliberation during testing. Specifically, this module stores the Key and Value pairs of multi-image input tokens obtained after each decoder layer in a dedicated memory bank. When decoding specific coordinates, the model retrieves corresponding visual tokens using image IDs and coordinates, then computes attention between the query vectors (derived from subgraph token sequences processed through each decoder layer) and the stored multi-image Key/Value pairs in the memory bank.

Furthermore, considering that existing works lack datasets meeting our requirements, we constructed a new dataset named **CMMCoT-260k**, which constitutes an innovative benchmark specifically designed for complex multi-image multimodal tasks, comprising 260,000 meticulously curated data instances. Distinct from conventional datasets, CMMCoT-260k's uniqueness resides in its incorporation of explicit reasoning chains within each instance. These structured reasoning pathways not only facilitate deep semantic parsing of hybrid text-image data but also integrate spatial coordinates and entity-specific imagery, thereby enabling hierarchical reasoning analysis.

Extensive experiments conducted on both multi-image and single-image benchmarks demonstrate the effectiveness of our model. Moreover, visualization experiments have also demonstrated that compared to the traditional GPT-4o and Qwen2.5-VL, our model not only improves the accuracy of responses but also significantly enhances the interpretability of the intermediate reasoning process, making it more accessible for human understanding.

Dataset Construction

Building upon multiple datasets including GRIT (Peng et al. 2023), Flickr30k-Entities (Plummer et al. 2015), VoCoT (Li et al. 2024d), and MANTIS (Jiang et al. 2024), we have constructed a complex multi-image, multi-modal Chain-of-Thought (CoT) dataset, referred to as the **CMMCoT-260K** dataset. It consists of 260,000 instances, encompassing four distinct task types: Caption, Co-reference, Comparison, and Reason.

For the Caption and Co-reference tasks, we employed a straightforward data integration pipeline. In contrast, for the more challenging Comparison and Reason tasks, we designed a more complex data processing pipeline that includes the following steps:

- **Constructing QA Rationale Chains:** We generate CoT annotations through a fully automated pipeline based on the methodology from (Zelikman et al. 2024). Initially, GPT-4o is used to generate a preliminary reasoning chain and answer from the question. If the answer is correct, the chain is retained. If incorrect, the question is paired with the pre-annotated correct answer and re-fed into GPT-4o to generate a refined rationale. This process uses answer correctness as a quality filter to ensure the reliability of the generated rationales.
- **Entity Extraction:** The Qwen3-235B-A22B model is utilized to extract textual entities from QA dialogues, providing a foundation for subsequent entity localization and relationship summarization.
- **Entity Detection:** To ensure localization accuracy, we adopt a two-stage validation mechanism. First, the Qwen-VL-max model generates initial bounding boxes for entities. Subsequently, GPT-4o calculates the Intersection over Union (IoU) for these boxes, and only samples with an IoU ≥ 0.9 are retained to ensure precision.
- **Entity Relationship Summarization:** When a single entity corresponds to multiple bounding boxes across images or detections, we apply spatial fusion. The smallest top-left and largest bottom-right coordinates from all relevant boxes are used to create a single, unified bounding box. This approach simplifies the CoT structure while maintaining complete entity coverage, enhancing the clarity and logical coherence of the reasoning chain.

In summary, through this pipeline, we have successfully transformed multi-image QA datasets into a comprehensive and powerful multi-image, multi-modal CMMCoT dataset, establishing an essential foundation for sophisticated multimodal reasoning systems.

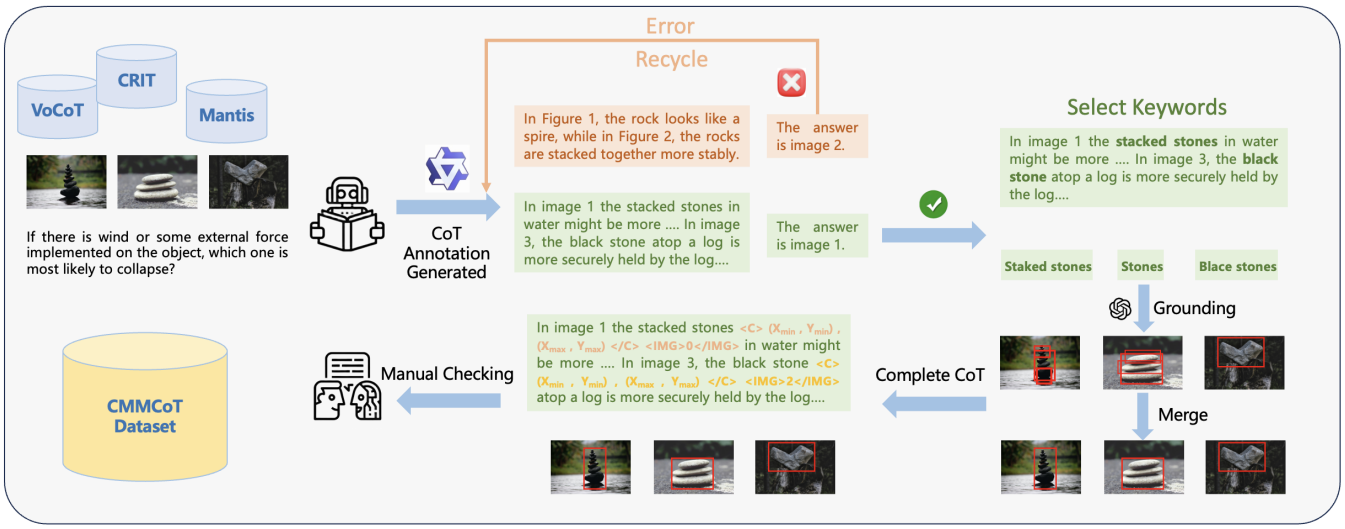

Figure 1: Data generation process of our proposed CMMCoT dataset. The construction of the CMMCoT dataset mainly concerns four parts: the generation of QA rationale chains, the extraction of textual entities, the detection and validation of visual entities, and the spatial fusion and summarization of entity groundings.

Methods

Baseline

Along with the proposed dataset, we also develop a novel multi-modal framework named CMMCoT, which employs Qwen2-VL (Wang et al. 2024a) as the baseline model. The primary reason for selecting Qwen2-VL is its trained Vision Transformer (ViT) that supports Naive Dynamic Resolution, enabling the processing of images at arbitrary resolutions and dynamically converting them into a variable number of visual tokens. Furthermore, Qwen2-VL introduces Multimodal Rotary Position Embedding (M-RoPE), which effectively models the positional information of multimodal inputs. This approach reduces the positional ID values of images and videos, allowing the model to extrapolate to longer sequences during inference.

Multimodal Sequence Representation

CMMCoT represents complex multi-image and textual data in an interleaved visual-textual format. For complex multi-image tasks, it is essential to compare and contrast the associations and differences among different images, between different entities within the same image, and between different entities across different images during the reasoning process. Therefore, we introduce special image index tokens to refer to specific input images, formatted as <IMG>0</IMG>. Here, <IMG> and </IMG> are special tokens indicating the start and end of an image index. This indexing logic can be extended to accommodate more images.

Additionally, we represent different entities using coordinates and visual tokens. The coordinate format is similar to <|box_start|>($<x_0>$, $<y_0>$) , ($<x_1>$, $<y_1>$)<|box_end|>, where <|box_start|> and <|box_end|> are special markers indicating the beginning and end of coordinate information. We use bounding boxes ($<x_0>$, $<y_0>$) and ($<x_1>$, $<y_1>$) as the entity coordinates, with x and y normalized relative to the image dimensions, ranging from 0 to 1000. The coordinate values are tokenized and embedded as textual data.

For visual information pertaining to multiple images and entities, we use special markers <|vision_start|> and <|vision_end|> to denote the start and end positions of visual content. During the training process, we obtain entity images based on their coordinates and image indices and encode each entity image using a visual encoder. When encoding the entity images through the visual encoder, we limit the minimum resolution of the entities to 512 pixels, which allows us to extract more detailed features from the entity images (Li et al. 2024c).

Training: To achieve better performance on complex multi-image tasks without compromising single-image performance, our CMMCoT method employs a two-stage training strategy:

- Stage 1: Multi-Image Training. We begin by training on our self-constructed CMMCoT-260k dataset, which is specifically designed for complex multi-image tasks. This phase aims to enable the model to handle intricate tasks involving multiple images.
- Stage 2: Mixed Training. In this phase, the CMMCoT-260k dataset is mixed with a general dataset at a 1:1 sampling ratio for training. The goal of this strategy is to significantly alleviate catastrophic forgetting caused by multi-image tasks while retaining the model's general visual understanding capabilities.

The training objective function is to minimize the loss in predicting the next token. During training, for the CMMCoT-260k dataset, the model input consists of multiple images and related questions, and the output is an answer

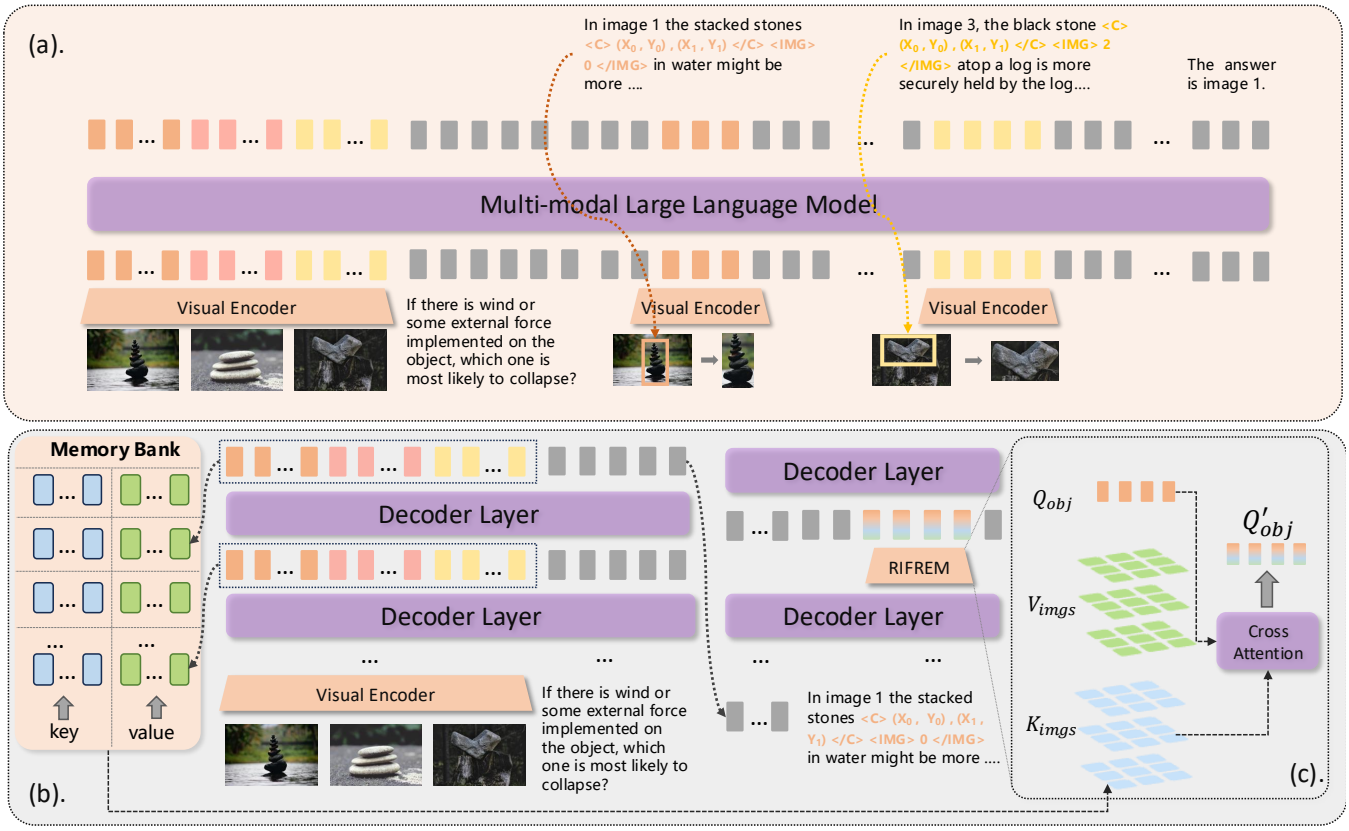

Figure 2: **Illustration of the overall framework of CMMCoT.** Part (a) depicts the training structure, with the input and output shown below and above the model, respectively. Part (b) presents the inference structure of the model, where the memory bank stores the K and V for each layer of the input images. The RIFREM module is integrated between different decoder layers during inference. Part (c) represents the detailed structure of the RIFREM module.

that encompasses the reasoning process, entity coordinates, and entity images. We append the prompt *"Please answer the question with reasoning and identify key objects."* after the question to enhance the model's reasoning ability. Entity images are extracted based on the provided entity coordinates and image indices and are encoded using a visual encoder. During training, the loss is calculated only on the text, coordinates, and special tokens; the entity image part does not contribute to the loss. For the general dataset, the training process follows the standard multimodal training procedure.

Inference: During the inference process, we also apply the prompt *"Please answer the question with reasoning and identify key objects."* to stimulate the model's reasoning capabilities in both single-image and multi-image tasks. The output of text and coordinates is consistent with that of standard multimodal models. However, when the </IMG> token is predicted, the entity image is extracted based on the image index and coordinates. Subsequently, the proposed RIFREM module is utilized to extract relevant features between the entity and the input multiple images, thereby enriching the entity features and continuing to reason over the subsequent tokens.

RIFREM

Previous studies (Snell et al. 2024; Prystawski, Li, and Goodman 2024) have demonstrated that expanding model computational capacity during inference can significantly enhance performance metrics. For multi-image tasks, a natural approach involves comparing critical visual regions with relevant areas in other images through joint reasoning when identifying pivotal visual attention zones.

To enable cross-image feature mining among entity-related images during inference, we propose the **Retrieval-based Image Feature Reasoning Enhancement Module (RIFREM).** This framework treats entity image features as queries (Q) while utilizing multi-image features as keys (K) and values (V) to retrieve relevant visual cues from input multi-image candidates. Specifically, we maintain a memory bank $\mathcal{M}$ to store the key-value pairs of multi-image sequences from each decoder layer during multi-image input processing. The memory bank can be formulated as:

$$\mathcal{M} = \{(K_l^{(i)}, V_l^{(i)})\}_{l=1,2,\ldots,L;i=1,2,\ldots,N} \quad (1)$$

where $K_l^{(i)}, V_l^{(i)}$ represents the keys and values from the l-th decoder layer of the i-th multi-image sequence, l is the number of decoder layers, and N is the number of multi-image sequences processed.

Model	Params	BLINK	Mantis	NLVR2	MVBench	Q-Bench	Avg
LLaVA-v1.5	7B	37.1	41.9	52.1	36.0	53.9	44.2
LLaVA-v1.6	7B	39.6	45.6	58.9	40.9	58.9	48.8
LLaVA-v1.5-MIA-DPO	7B	42.9	44.2	54.2	39.5	–	–
LLaVA-OV	7B	48.2	64.2	89.4	56.7	74.5	66.6
Mantis-Idefics2	8B	49.1	57.1	89.7	51.4	75.3	64.5
InternVL3	8B	55.5	70.1	88.5	75.4	75.9	73.1
Qwen2-VL	7B	51.8	64.8	84.0	50.5	71.2	64.5
Qwen2-VL (Ours)	7B	**52.3**	**70.9**	**88.7**	**51.3**	**72.2**	**67.1**
Qwen2.5-VL	7B	55.3	69.8	88.3	74.7	77.7	73.2
Qwen2.5-VL (Ours)	7B	**56.8**	**72.2**	**89.9**	**75.8**	**78.5**	**74.6**
InternVL3	2B	50.3	65.9	85.4	70.4	71.5	68.7
Qwen2-VL	2B	43.6	37.7	74.1	37.5	58.5	50.3
Qwen2-VL (Ours)	2B	**44.8**	**39.8**	**76.9**	**40.1**	**61.7**	**52.7**
Qwen2.5-VL	3B	49.1	62.7	86.2	71.3	74.9	68.8
Qwen2.5-VL (Ours)	3B	**51.4**	**68.5**	**88.9**	**73.1**	**75.2**	**71.4**

Table 1: Performance comparison of SOTA models on multi-image benchmarks. Rows in light gray background and darker gray background represent Qwen2-VL and Qwen2.5-VL variants, respectively. Best results per benchmark are in **bold**.

During CoT prediction, when encountering the </IMG> token in the reasoning process, we inject entity image tokens into decoder layers and extract their query vectors based on position IDs. These queries then engage in cross-attention with corresponding key-value pairs retrieved from the memory bank. The attention mechanism can be formally described as scaled dot-product attention:

$$Q' = \text{Att}(Q, K_{\mathcal{M}}, V_{\mathcal{M}}) = \text{softmax}\left(\frac{QK_{\mathcal{M}}^T}{\sqrt{d_k}}\right)V_{\mathcal{M}} \quad (2)$$

where $K_{\mathcal{M}}$ and $V_{\mathcal{M}}$ are the keys and values retrieved from the memory bank $\mathcal{M}$ corresponding to the current query Q, and d_k is the dimension of the keys. These refined Q' subsequently propagate through subsequent reasoning stages. Our ablation studies reveal that incorporating RIFREM modules across different decoder layers introduces non-trivial latency overheads, prompting systematic comparisons between performance gains and computational costs.

Experiments

Evaluation Benchmarks **NLVR2** (Suhr et al. 2018) evaluates visual-textual reasoning through cross-image logical analysis, presenting paired images with declarative statements for propositional verification. It employs constrained selection on the standard test-public partition. **Qbench** (Wu et al. 2023) assesses multimodal models' critical analysis of benchmarking parameters, focusing on visual perception in quality assessment tasks. Our experiment employs its Qbench2-A1-dev subset with multiple-choice comparative visual analysis. **Mantis-Eval** contains 217 multi-image analysis tasks across conceptual domains (dimensional assessment, mass estimation). Manually curated from web-sourced images, it combines closed and open-response formats evaluated on standardized test partitions. **BLINK** (Fu et al. 2025) tests rapid visual cognition (depth relationships, feature matching, digital forensics, spatial-temporal reasoning). It includes multi-image perceptual similarity tasks, measured through standardized validation protocols. **MVBench** (Li et al. 2024b) evaluates temporal reasoning in video comprehension through 20 tasks analyzing frame dynamics. Using 8-frame sampling per video, results are quantified via standardized test-set measurements. **Single Image Benchmarks:** We also conduct experiments on single-image tasks to further demonstrate the effectiveness of our proposed model: MMMU (Yue et al. 2024), MMStar (Chen et al. 2024), SQA (Lu et al. 2022), RealWorldQA, MME (Yin et al. 2023), POPE (Li et al. 2023), HallBench (Guan et al. 2023).

Experiments Setting We conduct supervised fine-tuning (SFT) based on the Qwen2.5-VL-7B architecture. The optimization process employs AdamW with $\beta = 0.95$ and a weight decay of 0.1, coupled with a cosine learning rate scheduler. The training framework utilizes the DeepSpeed ZeRO-3 strategy for efficient parameter optimization. The training regimen consists of two distinct phases: Stage 1: Initial learning rate of 1e-5 with 2 training epochs. Stage 2: Reduced learning rate of 1e-6 with 1 training epoch at batch size 256. This two-phase approach enables progressive refinement of model parameters, where the first stage establishes coarse-grained feature representations and the second stage performs fine-grained adjustment with reduced stochasticity through smaller learning rates and larger batch sizes.

Models	Params	MMMU	MMStar	SQA	RealWorldQA	MME	POPE	HallBench	Avg
LLaVA-v1.5	7B	35.1	32.9	66.6	48.9	58.4	78.0	35.8	50.8
LLaVA-v1.6	7B	35.8	37.6	69.7	60.8	63.6	86.1	30.3	54.8
LLaVA-MIA-DPO	7B	36.3	32.9	67.6	–	–	87.2	–	–
LLaVA-OV	7B	48.8	61.7	94.8	66.3	71.4	87.1	48.3	68.4
Mantis-Idefics2	8B	40.1	46.4	77.3	59.3	64.5	82.8	56.7	61.0
InternVL3	8B	62.7	68.7	97.9	71.4	86.5	90.4	49.0	75.2
Qwen2-VL	7B	**51.6**	57.7	85.5	63.3	**82.7**	**88.4**	50.6	68.5
Qwen2-VL (Ours)	7B	48.2	**58.9**	**88.9**	**65.0**	81.5	86.3	**60.5**	**69.9**
Qwen2.5-VL	7B	**58.6**	63.9	89.0	68.4	82.6	**85.9**	51.9	71.4
Qwen2.5-VL (Ours)	7B	57.5	**66.4**	**96.8**	**71.6**	**83.5**	89.2	**63.6**	**75.5**
InternVL3	2B	48.6	61.1	95.8	65.1	78.1	**90.1**	41.9	68.7
Qwen2-VL	2B	**41.1**	48.0	73.2	62.3	**66.8**	**85.8**	41.7	59.8
Qwen2-VL (Ours)	2B	39.4	**49.2**	**78.6**	**62.6**	65.7	84.0	**52.5**	**61.7**
Qwen2.5-VL	3B	**53.1**	55.8	81.4	65.5	78.6	85.9	46.6	66.7
Qwen2.5-VL (Ours)	3B	53.0	**57.3**	**88.6**	**67.2**	**79.8**	**87.1**	**48.6**	**68.8**

Table 2: Performance Comparison of SOTA Models across Single-image Benchmarks. The Qwen2-VL series are highlighted with light gray background and Qwen2.5-VL series with darker gray background, while other models use white background. Bold numbers indicate the best performance for each benchmark.

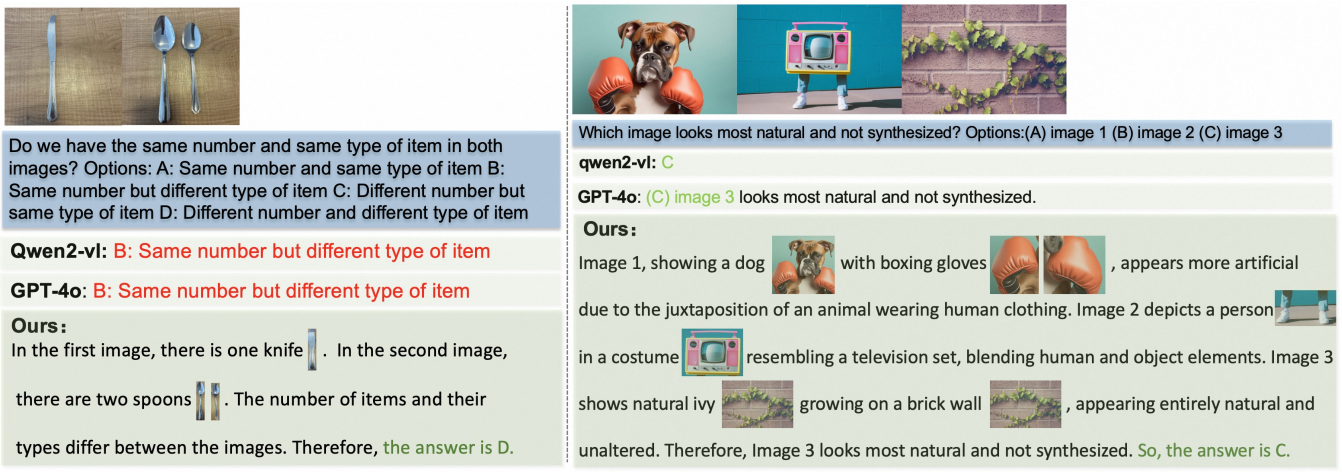

Figure 3: Visualization results of our CMMCoT task to illustrate the difference between our model and previous methods.

Compared Methods To evaluate and compare the effectiveness of multi-image models, we conducted comprehensive benchmarking on several state-of-the-art architectures: LLaVA-v1.5 (Liu et al. 2024a), LLaVA-v1.6 (Li et al. 2024a), LLaVA-v1.5-MIA-DPO (Liu et al. 2024c), Mantis-Idefics2 (Jiang et al. 2024), LLaVA-OV (Li et al. 2024a), Qwen2-VL (Wang et al. 2024b), InternVL3 (Zhu et al. 2025) and Qwen2.5-VL (Bai et al. 2025). Our evaluation framework simultaneously addresses both multi-image and single-image tasks, enabling comparative analysis of model performance across different input configurations.

Performance Evaluation Multi-image performance: The experimental results across multiple diagnostic benchmarks demonstrate the consistent superiority of our model across both 7B and 3B parameter configurations. As shown in Tab. 1, our 7B variant achieves state-of-the-art performance on Mantis-Eval (72.2), MVBench(75.8), Q-Bench(78.5), and competitive results on NLVR2 (89.9) and BLINK (52.3), outperforming comparable 7B models like Qwen2.5-VL by 1.4 average points. Notably, the scaled-down 3B version maintains robust capabilities, surpassing its parameter-matched counterpart Qwen2.5-VL-3B by 2.6 average points while exhibiting particular strengths in temporal reasoning (MVBench: 73.1) and visual-textual verification (NLVR2: 88.9). This dual-scale effectiveness suggests our architecture successfully balances model capacity with cross-modal alignment efficiency, preserving critical visual reasoning abilities even in parameter-constrained scenarios. The performance parity across distinct evaluation paradigms from multi-image logical analysis (Mantis) to perceptual similarity judgments (BLINK) validates our

Grounding	MCoT	RIFREM	BLINK	Mantis	NLVR2	MVBench	Q-Bench	Avg
✘	✘	✘	55.3	69.8	88.3	74.7	77.7	73.2
✓	✘	✘	55.2	70.4	88.7	74.5	77.9	73.3
✓	✓	✘	**57.1**	71.6	89.4	75.4	**78.7**	74.4
✓	✓	✓	**56.8**	**72.2**	**89.9**	**75.8**	78.5	**74.6**

Table 3: Performance (%) of Qwen2.5-VL-7B variants across benchmarks with different module combinations

approach's adaptability in handling diverse visual reasoning tasks through enhanced feature interaction mechanisms.

Single-image performance: Similarly, as shown in Tab. 2 in single-image datasets, our model consistently achieves superior performance across different scales compared to prior architectures, demonstrating that the proposed multi-modal hybrid sequential chain reasoning strategy effectively enhances comprehension capabilities for single-image tasks. This cross-scale effectiveness confirms the generalizability of our methodology beyond multi-image analytical scenarios, particularly in improving fundamental visual-semantic alignment through structured reasoning pathways.

Quantitive Analysis

To further evaluate how multimodal hybrid sequence CoT enhances model understanding in complex multi-image scenarios, we visualize representative examples in Fig. 3. The left example in Fig. 3 highlights our model's superiority in object localization and quantitative reasoning. While other models fail to correctly detect or count cutlery items (e.g., knives and spoons), our model accurately identifies their positions and quantities, demonstrating robust counting ability even in ambiguous visual contexts. The right example illustrates our model's advantage in anomaly detection and logical reasoning. Although alternative models may provide correct answers, our method not only identifies commonsense violations in the visual input but also offers explicit causal interpretations. This capacity for contextual inconsistency detection underscores the enhanced reasoning capabilities enabled by our approach.

Ablation Study

Impact of Different Modules. We conducted a series of ablation studies to examine the contributions of individual components, as shown in Table 3. We first analyzed the effect of using only grounding-based information during training. This setting yielded a modest 0.1-point improvement in average score, suggesting limited benefit when used in isolation. Next, we investigated the impact of combining grounding with entity images. This joint supervision led to a notable performance gain, improving the average score from 73.2 to 74.4. The result highlights the synergy between grounding cues and visual features in enhancing multimodal representation learning. We further evaluated the RIFREM module applied during inference. Its integration provided an additional boost, raising the average score to 74.6 by exploiting cross-image feature relationships. Together, these experiments clarify the individual and combined contributions

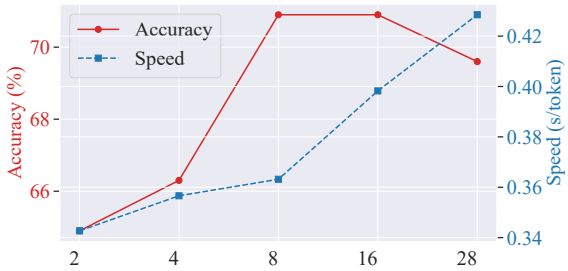

Figure 4: **Ablation study on the number of RIFREM module layers:** The red line represents the performance effects, while the blue line indicates the impact on latency.

of each module. They underscore the importance of coordinated feature utilization—across both training and inference—for optimizing performance in multimodal tasks.

Impacts of RIFREM. We evaluate the effect of integrating the RIFREM module at different network depths during inference, balancing performance gains and latency overhead (Figure 4). Five configurations were tested, from shallow (Group 1: layers 0 and 27) to full-depth integration (Group 5: all layers). Results show that applying RIFREM only at the first and last layers degrades performance, likely due to disrupted information flow. In contrast, progressive integration across the network (Groups 2–5) enables iterative refinement and better feature consistency. Group 3 (8 evenly distributed layers) achieves the best trade-off between accuracy and efficiency, and is used in all subsequent experiments.

Conclusion

In this paper, we propose CMMCoT, a framework enabling systematic slow-thinking for multi-modal large language models (MLLMs) in complex multi-image scenarios. By integrating coordinate-guided visual token extraction and a Retrieval-based Image Feature Reasoning Enhancement Module (RIFREM), CMMCoT mitigates error accumulation and enhances cross-image visual concept tracking. Evaluations across six benchmarks demonstrate state-of-the-art performance, with RIFREM facilitating dynamic deliberation over multi-image contexts during inference. Case studies highlight improved accuracy and interpretability over GPT-4V. This work underscores the necessity of structured reasoning mechanisms for advancing MLLMs, paving the way for future research in dynamic multi-modal reasoning and memory-efficient architectures.

Acknowledgments

This research was supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant LD24F020016, the Key R&D Program of Zhejiang Province 2025C01075, 2023C01043, the National Natural Science Foundation of China under Grant 62576313, and Alibaba Group through Alibaba Research Intern Program.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *ArXiv*, abs/2502.13923.

Brown, B.; Juravsky, J.; Ehrlich, R.; Clark, R.; Le, Q. V.; Ré, C.; and Mirhoseini, A. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. 2024. Are We on the Right Way for Evaluating Large Vision-Language Models? *arXiv preprint arXiv:2403.20330*.

Du, Y.; Liu, Z.; Li, Y.; Zhao, W. X.; Huo, Y.; Wang, B.; Chen, W.; Liu, Z.; Wang, Z.; and Wen, J.-R. 2025. Virgo: A Preliminary Exploration on Reproducing o1-like MLLM. *arXiv preprint arXiv:2501.01904*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*.

Feng, K.; Ma, Y.; Wang, B.; Qi, C.; Chen, H.; Chen, Q.; and Wang, Z. 2025. Dit4edit: Diffusion transformer for image editing. In *AAAI*, volume 39, 2969–2977.

Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N. A.; Ma, W.-C.; and Krishna, R. 2025. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 148–166. Springer.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. 2023. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. *arXiv preprint arXiv:2310.14566*.

He, W.; Fu, S.; Liu, M.; Wang, X.; Xiao, W.; Shu, F.; Wang, Y.; Zhang, L.; Yu, Z.; Li, H.; et al. 2025. Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. In *AAAI*, volume 39, 17123–17131.

Jiang, D.; He, X.; Zeng, H.; Wei, C.; Ku, M.; Liu, Q.; and Chen, W. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

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*.

Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Li, Y.; Zhang, Y.; Wang, C.; Zhong, Z.; Chen, Y.; Chu, R.; Liu, S.; and Jia, J. 2024c. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Li, Z.; Luo, R.; Zhang, J.; Qiu, M.; and Wei, Z. 2024d. VoCoT: Unleashing Visually Grounded Multi-Step Reasoning in Large Multi-Modal Models. *arXiv preprint arXiv:2405.16919*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, M.; She, D.; Pang, J.; Huang, Q.; Ying, J.; He, W.; Hou, Y.; and Fu, S. 2025. TFCustom: Customized Image Generation with Time-Aware Frequency Feature Guidance. In *CVPR*, 2714–2723.

Liu, Z.; Zang, Y.; Dong, X.; Zhang, P.; Cao, Y.; Duan, H.; He, C.; Xiong, Y.; Lin, D.; and Wang, J. 2024c. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. *arXiv preprint arXiv:2410.17637*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

Ma, Y.; Feng, K.; Zhang, X.; Liu, H.; Zhang, D. J.; Xing, J.; Zhang, Y.; Yang, A.; Wang, Z.; and Chen, Q. 2025. Follow-Your-Creation: Empowering 4D Creation through Video Inpainting. *arXiv preprint arXiv:2506.04590*.

Ma, Y.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; Liu, W.; et al. 2024. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia*, 1–12.

Ni, M.; Fan, Y.; Zhang, L.; and Zuo, W. 2024. Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning. *arXiv preprint arXiv:2410.03321*.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, 2641–2649.

Prystawski, B.; Li, M.; and Goodman, N. 2024. Why think step by step? Reasoning emerges from the locality of experience. *Advances in Neural Information Processing Systems*, 36.

Shao, H.; Qian, S.; Xiao, H.; Song, G.; Zong, Z.; Wang, L.; Liu, Y.; and Li, H. 2024. Visual cot: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Snell, C.; Lee, J.; Xu, K.; and Kumar, A. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.-Y.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024b. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *ArXiv*, abs/2409.12191.

Wang, W.; Dong, L.; Cheng, H.; Liu, X.; Yan, X.; Gao, J.; and Wei, F. 2024c. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Wang, Y.; Liu, M.; He, W.; Zhang, L.; Huang, Z.; Zhang, G.; Shu, F.; Tao, Z.; She, D.; Yu, Z.; et al. 2025. Mint: Multimodal chain of thought in unified generative models for enhanced image generation. *arXiv preprint arXiv:2503.01298*.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Li, C.; Sun, W.; Yan, Q.; Zhai, G.; et al. 2023. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*.

Xu, G.; Jin, P.; Hao, L.; Song, Y.; Sun, L.; and Yuan, L. 2024. LLaVA-o1: Let Vision Language Models Reason Step-by-Step. *arXiv preprint arXiv:2411.10440*.

Yang, J.; Zhang, H.; Li, F.; Zou, X.; Li, C.; and Gao, J. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.

Zelikman, E.; Wu, Y.; Mu, J.; and Goodman, N. D. 2024. STaR: Self-taught reasoner bootstrapping reasoning with reasoning. In *Proc. the 36th International Conference on Neural Information Processing Systems*, volume 1126.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.