DCoT: Dual Chain-of-Thought Prompting for Large Multimodal Models

Zixi Jia[*] jiazixi@mail.neu.edu.cn
Jiqiang Liu 2302152@stu.neu.edu.cn
Hexiao Li lihexiao2021@163.com
Qinghua Liu 2302155@stu.neu.edu.cn
Hongbin Gao ghb2323@163.com
Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Inference augmentation techniques such as Chain-of-Thought have already made their mark in Large Language Models (LLMs). However, transferring these advances to Large Multimodal Models (LMMs) presents greater challenges. Drawing inspiration from human cognitive processes, this paper proposes a plug-and-play Dual Chain-of-Thought strategy, a novel pipeline that combines visual and textual guidance to improve the performance of LMMs in complex multimodal tasks. The DCoT strategy uses a dual guidance mechanism to use bounding box markers to guide the model's attention to the image region related to the query problem in the visual aspect, so as to achieve fine-grained image guidance, and in the text aspect, we propose a Fast In-Context Retrieval Framework (FICRF) dynamically and automatically obtains the most suitable examples from the well-built demonstration example cluster as context guidance according to the current problem. This bimodal approach that utilizes visual and textual guidance enhances the inference capabilities of LMMs. Extensive experiments on different LMMs and benchmark datasets have validated its effectiveness, opening up a new path in multimodal inference. Showcasing how the synergistic combination of visual and textual instructions can take the performance of these models to new heights, while demonstrating the potential of Chain-of-Thought and In-Context Learning as a superior alternative to the fine-tuning of LMMs.

Keywords: Large Multimodal Models, Multimodal Chain-of-Thought, In-Context Learning

1. Introduction

With the rapid rise of Large Language Models (LLMs) such as GPT-4 Achiam et al. (2023) and GLM-4 Zeng et al. (2024) heralding a new era of artificial intelligence, they have demonstrated impressive capabilities in multiple fields and opened the way to higher forms of intelligence. Recognize that language, while the fruit of human ingenuity, does not exist in isolation. The real world is a multi-sensory, multi-dimensional interweaving of information, and language models alone cannot fully capture all the information in the real situation Roh et al. (2022). Therefore, the leap from LLMs to Large Multimodal Models (LMMs) is an inevitable trend in the development of artificial intelligence, which prompts researchers to develop LMMs. Recent works, such as LLaVA Liu et al. (2024b), Blip2 Li

[*] Corresponding author

et al. (2023a), and Qwen VL Bai et al. (2023), are excellent exploratory works for LMMs that utilize visual encoders and LLMs to train on large-scale image-text datasets to align the output of the vision model with the input space of LLMs. However, frozen visual encoders are often trained on low-resolution images, resulting in relatively weak fine-grained image understanding of LMMs, which necessitates adjustments to improve their usefulness Li et al. (2024b). Despite effective improvements, fine-tuning methods require high-quality instruction datasets and independent training of each model, but the sheer size of these models, often measured in billions or even hundreds of billions of parameters, makes fine-tuning them computationally demanding and costly. In addition, fine-tuning can lead to a decrease in generalization ability Zhai et al. (2024), severe hallucinatory phenomena Li et al. (2023b). And since most of the state-of-the-art LMMs are closed-source, it is not feasible to apply fine-tuning to these.

An emerging trend in AI research aims to unlock the potential of large multimodal models with cost-effective solutions. A key finding in this paradigm shift is the emergence of remarkable capabilities Chain-of-Thought Wei et al. (2022) and In-Context Learning Coda-Forno et al. (2023), as the parameter size of LMMs increases. These inference augmentation techniques, which do not require additional training, have attracted a lot of attention because of their plug-and-play nature, which can enhance the performance of the task without updating the pre-training parameters, and provide a simple and efficient way to customize LMMs. The Chain-of-Thought approach enhances the interpretability of the model's output by simulating the step-by-step human thought process. Existing CoT prompts can be divided into two paradigms: Zero-shot-CoT and Few-shot-CoT. Zero-shot-CoT directly leverages a single prompt like "Let's think step by step" Kojima et al. (2022) to generate a chain of inference. For example, DDCoT Zheng et al. (2023) uses negative spatial cues to explicitly point out uncertainties in the process of generating reasons for decomposing sub-problems, reducing false reasoning or hallucinations. VoCoT Li et al. (2024a) visually represents object concepts in a multimodal crossover and alignment manner using instruction optimized datasets, effectively bridging the modal differences of LMMs in long text processes. Few-shot-CoT uses some inference examples as hints. In-Context Learning by embedding examples in the input, enables the model to infer and generate appropriate responses based on the examples, demonstrating the amazing generalization power inherent in language models. Therefore, we believe that for the future of the large models domain, In-Context Learning with appropriate prompts may be a better solution than fine-tuning.

This study aims to explore ways to overcome the above challenges and develop a plug-and-play method to improve LMMs' inference performance on fine-grained image comprehension and complex question answering. To achieve this, we propose a Dual Chain-of-Thought strategy that combines visual and textual guidance, which is a training-free guidance method. DCoT strategy inspired by the way humans process signals, by mimicking human vision and their thinking. For the image part, our strategy focuses on fine-grained guidance, explicitly based on a specific image area, focusing on the important parts of the image related to the question, reducing the interference of redundant information. For the text part, We introduced In-Context Learning to retrieve the most similar questions to the relevant question categories among the 80 examples of 8 different question types that we carefully constructed as prompts to help the model better understand the questions to answer. To the best of our knowledge, this is the first work to merge image and text guidance, and our contributions are as follows:

- We propose DCoT, a Chain-of-Thought method that combines image processing and text guidance, which is training-free and plug-and-play, providing valuable insights for future multimodal tasks.
- In the text guidance stage, we propose a Fast In-Context Retrieval Framework (FI-CRF), which improves the retrieval efficiency, and can be dynamically and adaptively retrieval according to the current questions.
- We conducted extensive experiments using multiple benchmark datasets on a series of LMMs to verify the effectiveness of our method.

2. Related Work

2.1. Large Multimodal Models

The success of Large Language Models has laid the foundation for the development of Large Multimodal models, which provide inference capabilities and a rich knowledge base for visual tasks by integrating LLMs and visual encoders. LLaVA Liu et al. (2024b) uses MLP as a visual-linguistic connector to map image features to the word embedding space of a pre-trained LLM. BLIP-2 Li et al. (2023a) uses Q-Former, which uses a learnable query to extract visual features from a frozen image encoder and align those features with the language model. Qwen-VL Bai et al. (2023) introduces a location-aware adapter to compress image features while retaining position information, which realizes the accurate alignment of image features with the LLMs input space. MiniGPT4 Zhu et al. (2023) is trained on visual language teaching data in the form of image subtitles. VistaLLM Pramanick et al. (2024) proposes a unified framework for LMMs with single and multiple visual scene inputs, and introduces an adaptive sampling algorithm to refine the NLP format mask of LMMs output.

2.2. Multimodal In-Context Learning

Multimodal In-Context Learning essentially uses images and their task descriptions to guide the model to generate more consistent answers. In-Context Learning is a new paradigm that improves performance on downstream tasks without updating any parameters and additional computational resources. Specifically, In-Context Learning utilizes the characteristics of language models that enable learning in specific contexts, absorbing and adapting new information from the context to output expected responses. However, this effectiveness depends on the appropriate choice of contextual examples.

Flamingo Alayrac et al. (2022) trained on a large-scale multimodal web corpus containing arbitrary interleaved text and images, showing that LMMs can also acquire such capabilities, and Emu2 Sun et al. (2024) achieved superior capabilities in multimodal In-Context Learning by training with unified autoregressive goals. Liu et al. (2023a) Multimodal In-Context Learning in the form of image subtitles using heterogeneous data retrievers. Luo et al. (2024) proposed a new multimodal unsupervised searcher MSIER to select context samples, which have achieved remarkable results in multimodal tasks.

2.3. Multimodal Chain-of-Thought

In the past, Multimodal Chain-of-Thought usually embeds visual information into the text explicitly, for example, Wu et al. (2023) uses the strategy of describing first and then deciding to obtain image information, and CCoT Mitra et al. (2024) describes the objects, attributes, and relationships in the image in json format in the form of scene graphs. However, a significant challenge with this approach is the loss of image detail when describing using natural language. Natural language is less precise than visual data when conveying complex visual information, such as subtle shifts in color and light in natural landscapes and artistic works, or the nuanced expressions in portraits, which are inherently challenging to articulate due to linguistic bottlenecks.

Existing work uses methods such as magnification Cao et al. (2024), cropping Liu et al. (2024c), masking Wan et al. (2024), and search localization Wu and Xie (2024) to focus on the local information of the image, because highlighting particularly relevant regions of the image can improve the performance of the model on various visual tasks by guiding the model to focus more closely on these regions of interest. Despite its achievements, LMMs' performance in visual question answering tasks is still limited by its lack of context and detail capture of complex visual scenes. Therefore, We propose a Dual Chain-of-Thought based on image and text, and our method not only captures fine-grained image features, but also provides textual guidance through contextual examples.

3. Method

In this section, we will first briefly introduce some preliminary knowledge about visual question answering tasks and LMMs, and then elaborate on the framework and implementation of our DCoT in detail.

3.1. Preliminaries

Generally speaking, for a visual question answering task T, given an image I and a related question Q in task T, the purpose of LMMs is to capture the information of the input image I and question Q to predict the answer A. This task requires not only the ability of the system to understand complex visual information, but also the flexibility and depth of natural language, which is highly dependent on the model's ability to integrate cross-modal information.

Therefore, the model first needs to understand and relate the content of the image modality and the question of the text modality interactively, and we need to find a mapping function from the image $i \in I$, the question $q \in Q$ to the answer $a \in A$, described as:

$$f(t) : (I, Q) \to A$$

Predict the best answer $a \in A$, and complete the conversion of multimodal input to textual answer. In general, this process uses a visual encoder $v_\phi(\cdot)$ (parameterized by ϕ) to efficiently encode image I to obtain the corresponding feature F_x, and then project the other modal feature F_x into the embedding space shared with question Q through the projector to obtain the aligned feature P_x.

Finally, use the elaborate prompt P to elicit the desired response from the LMMs. The LMMs thus generates a response as follows:

$$\hat{y}_a = \arg\max_a P\left(y \mid \mathrm{P}(c, i, q); \theta^{LMMs}\right)$$

where $\mathrm{P}(\cdot)$ denotes the combination of the prompt template c with the chain of thought and the input (i, q) as the input of the language model using a template that conforms to the θ^{LMMs} parameter heuristic.

3.2. Overall framework of DCoT

An overview of our approach is shown in Figure 1, which is a strategy formed by two-turn dialogue pipelines, the primary objective of the initial dialogue round is to concurrently analyze the original image and question input through a dual-pronged approach, obtaining the fine-grained image and ascertaining the categorical domain to which the posed query pertains, respectively, and before the second round of dialogue, the FICRF framework is used to retrieve the topK optimal examples as the context, and the time and computing resources consumed are almost negligible due to the retrieval in a fixed category and a small range. The purpose of the second round of dialogue is to obtain the final answer, and the fine-grained images obtained from the first round of dialogue and the topk context examples obtained by FICRF are used as new inputs for the LMMs to obtain the final response. DCoT consists primarily of visual guidance utilizing bounding box prompts and contextual text guidance implemented by a fast in-context retrieval framework, both of which are described in detail in Sections 3.3 and 3.4. The overall algorithm framework of DCoT is simplified to Algorithm 1. Among them, BBC represents the bounding box coordinates, Q_c represents the problem category, FGI represents fine-grained images, C_i represents the examples in the current problem category C, and $TopK$ represents the K most similar contextual examples finally retrieved.

It is worth noting that, since virtually all LMMs, such as Llava and Qwen-vl, acquire robust instruction-following and visual grounding capabilities upon training with large-scale image-text datasets, our proposed method does not necessitate additional training efforts.

Algorithm 1 Dual Chain-of-Thought

```
Input  : question $Q$, image $I$, Image Prompt $IP$, Question Classification Prompt $QCP$,
         Text Prompt $TP$
Output: Final Answer $FA$
$BBC, Q_C \leftarrow \mathrm{LMM}(Q, I, IP, QCP)$  // Generate coordinates and question category
$FGI \leftarrow \mathrm{draw\_rectangle}(I, BBC)$  // Generate Fine-Grained Image
for each $C_i$ in $Q_C$  // Retrieve top-K similar examples using FICRF
    $TopK \leftarrow Sim(Q, C_i)$ do
end
$FA \leftarrow \mathrm{LMM}(Q, FGI, \mathrm{TopK}, TP)$  // Generate final answer using DCoT
return $FA$
```

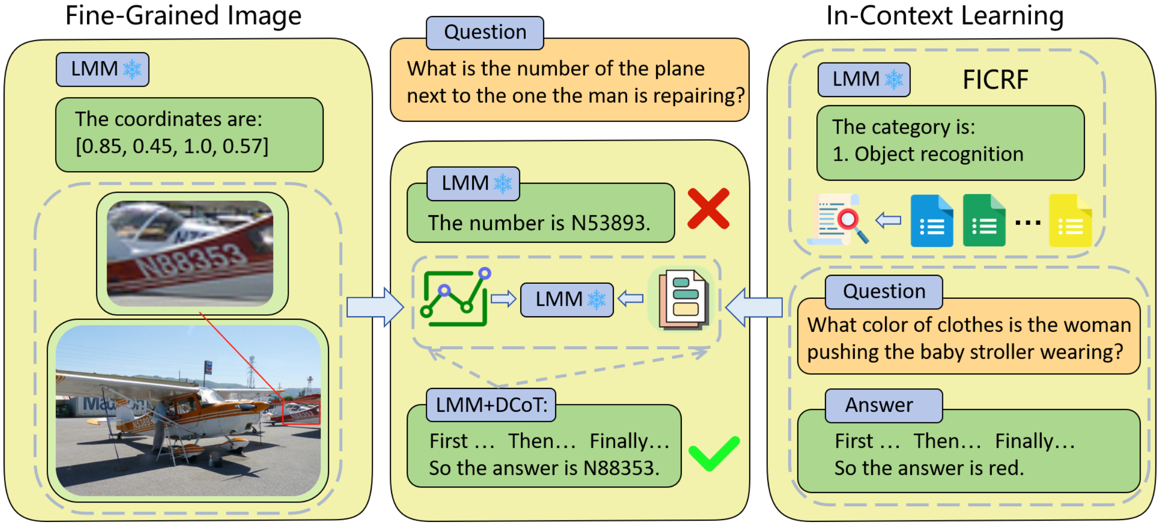
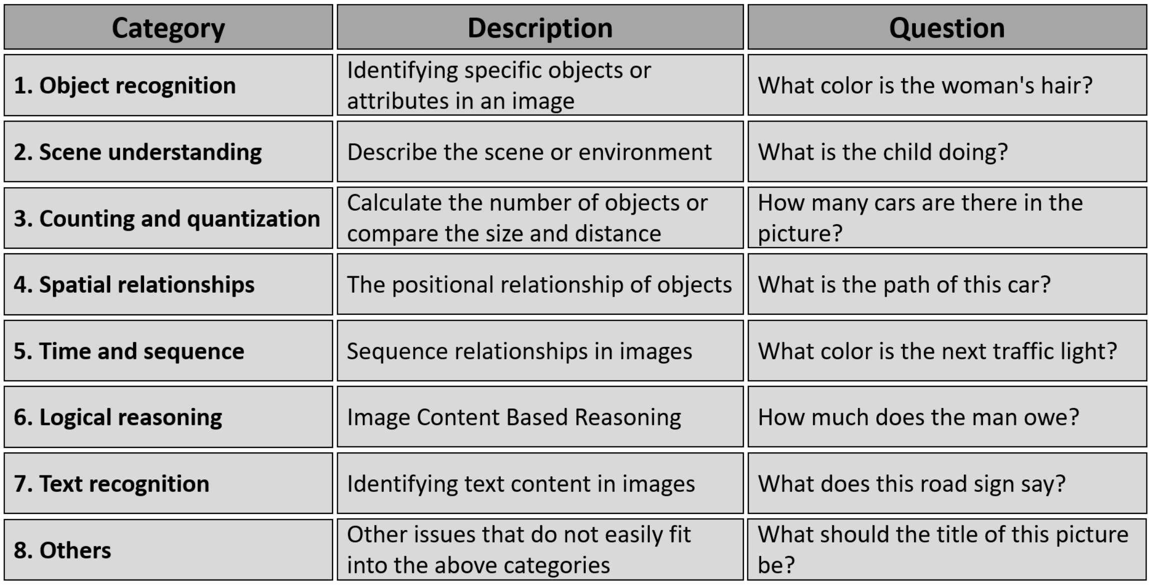


Figure 1: An overview of our Dual Chain-of-Thought (DCoT) method. Our DCoT method consists of two steps: (1) Firstly, obtaining fine-grained images through bounding box prompts. (2) Subsequently, obtain the k most relevant demonstration examples as context to generate responses.

3.3. Fine-Grained Images

For the image processing component of the input data, we employ a fine-grained guidance strategy. Central to this methodology is the directive for the model to concentrate on the most salient portions of the image pertinent to the question, achieved by explicitly delineating key areas within the image. We complete the acquisition of fine-grained images in the form of dialogues, specifically, we first add bounding box prompts to the question, which is a clear signal that the model is required to identify and focus on specific areas of the image that are directly related to the question. Specifically, we employ the instruction "*Please analyze the provided image and focus on the specific area related to the question. Generate bounding box coordinates:[x1, y1, x2, y2]*", where x1, y1 represents the coordinate coordinates of the upper left corner of the bounding box in the image, x2, y2 represents the coordinates of the upper right corner, and the value range is 0 to 1, representing the proportion of the bounding box occupying the width and height of the image, based on the left and upper boundaries of the original image. These coordinates are represented as numerical values embedded in natural language, with no additional formatting or special tokens, to maintain the convenience of natural interaction with humans.

The region coordinates are extracted from the response output of the first round of dialogue via a regular expression, and then marked with a striking red box on the original image as the picture input for the second round of dialogue. It's important to note that we didn't use images that only contain specific regions as input for the second round of dialogue, because we found that this approach relies heavily on the accuracy of the bounding box, and some questions do not have a specific area corresponding to them, which will lose a



Category	Description	Question
1. Object recognition	Identifying specific objects or attributes in an image	What color is the woman's hair?
2. Scene understanding	Describe the scene or environment	What is the child doing?
3. Counting and quantization	Calculate the number of objects or compare the size and distance	How many cars are there in the picture?
4. Spatial relationships	The positional relationship of objects	What is the path of this car?
5. Time and sequence	Sequence relationships in images	What color is the next traffic light?
6. Logical reasoning	Image Content Based Reasoning	How much does the man owe?
7. Text recognition	Identifying text content in images	What does this road sign say?
8. Others	Other issues that do not easily fit into the above categories	What should the title of this picture be?

Figure 2: Carefully constructed contextual examples, including 8 problem categories and their descriptions and questions.

large part of the information from the original image and cause the model to not be able to the correct response.

We use the LMMs own visual grounding capability to identify critical areas in the image and generate bounding boxes. This fine-grained guidance method ensures that the model can quickly locate the specific object or scene to which the question refers to, improving the accuracy of the model's understanding and analysis of the image content, resulting in more accurate and targeted responses.

3.4. In-Context Learning

In order to further improve the logical reasoning ability of the model, we introduce contextual learning strategy Kossen et al. (2024). We use cosine similarity to retrieve, with the aim of augmenting the CoT inference process with more relevant demo examples on multimodal tasks. To this end, we used the GPT4-Turbo 128k with an ultra-long context window to randomly select a large number of question and answer pairs from different datasets for coarse-grained summary and example generation of question categories, and then we manually carried out fine-grained secondary construction, and finally obtained 8 question types, 10 examples of each type, a total of 80 examples. As shown in Figure 2.

These Q&A pairs, while unrelated to the image content of the current query, have the answer (*"First... Then... Finally... So the answer is"*) implicitly embeds logical rules for the output, helping the model better understand and exploit the logical relationships in the context. In this way, our model is able to identify the key points of the problem to generate better answers. Compared with the method of using only sentence templates as supervised signals, our contextual learning strategy is able to provide rich logical clues for the output of the model.

Our goal is to automatically select the most appropriate example for the current issue Q from a well-constructed example. To this end, we propose a Fast In-Context Retrieval Framework (FICRF). Specifically, in the first round of conversations, we responded by prompting "*There are now eight question categories that are: ... Please determine which category the problem belongs to*", then measure the similarity between the feature vectors of the current query problem x_q and the feature vectors between the source domain c, and calculate the similarity between the current problem Q and the category question Q'

$$S(Q, Q'_c) = Sim(V(Q), V(Q'_c))$$

Among them, The dataset S includes $\mathbf{S}^c = \{(x_j^c, y_j^c)\}_{j=1}^{n_c}$ where x represents the feature vector and y represents the corresponding answer.

$$sim(x_q, x_j^c) = \frac{z(x_q) \cdot z(x_j^c)}{\|z(x_q)\| \left\|z(x_j^c)\right\|}$$

$z(x_j^c)$ refers to the feature vector extracted by the encoder Recall K of the most relevant contextual instances in the current question category c. When $K = 1$, we choose the optimal example pair as the prompt, $P = \{x^*, y^*\}$.

$$x^* = \arg\max_{x_n \in D} \sin(x_n, q_n)$$

When $K > 1$, we sort the retrieved examples based on the score and select the top-k example pair.

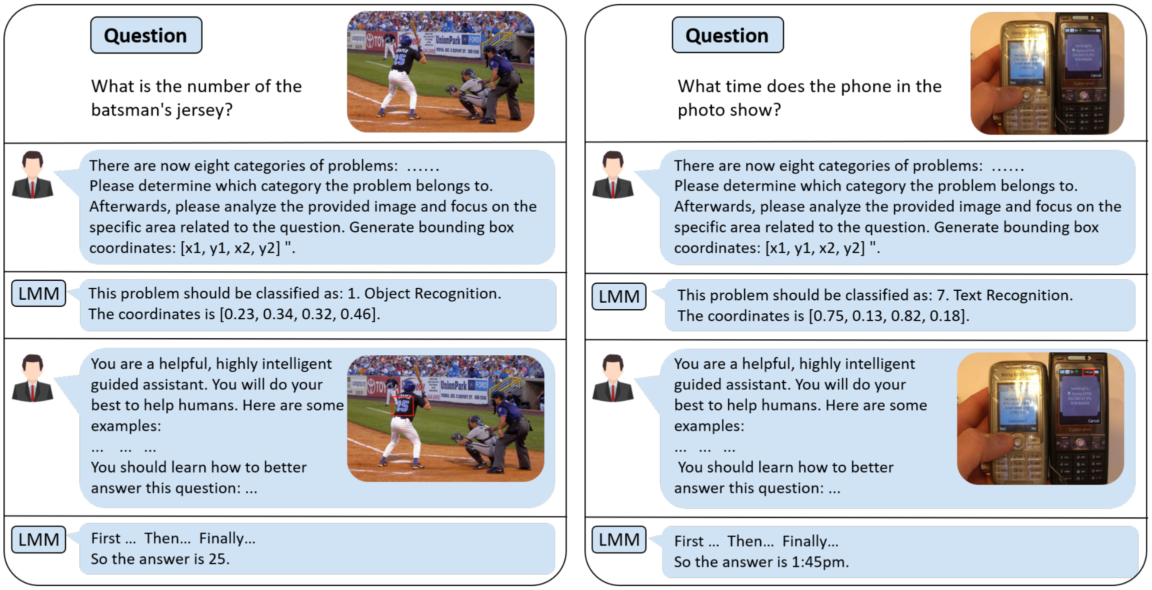
$$top_K(\{sim(x_q, x_j^i) : i = 1, \ldots, M; j = 1, \ldots n_i\})$$

FICRF determines the category to which the problem belongs first, rather than directly searching for all the examples, which greatly reduces the time complexity. The selected questions and their corresponding logical reasoning answers serve as contextual examples of LMMs, designed to enhance their understanding and performance of similar tasks without further training Zhang et al. (2024).

The specific case study is shown in Figure 3, for the question "What is the batsman's jersey number?", the first round of dialogue answers "1.Object Recognition" and the bounding box coordinates "[0.23, 0.34, 0.32, 0.46]" with a well-designed prompt guided model, then uses a Python script to mark the rectangular boxes and retrieve the examples. In the second round of dialogue, the image-guided and text-guided results serve as new inputs to the LMMs, and finally a logical output can be obtained: "First, analyze the question, mention the batter, jersey, and number. Then, we need to identify the characteristics of the batsman. In the image, wear a black top, white pants and blue socks, wearing a helmet and holding a bat. Finally, We need to find the number of the batsman's jersey, based on the fine image marking, although slightly obscured, it is not difficult to see that this is the number 25, so the answer is 25."

4. Experimental Results

In this section, we have conducted extensive experiments on a range of benchmark datasets using three types of LMMs, and in the following sections, we first describe these involved



What is the number of the batsman's jersey?

There are now eight categories of problems: ......
Please determine which category the problem belongs to.
Afterwards, please analyze the provided image and focus on the specific area related to the question. Generate bounding box coordinates: [x1, y1, x2, y2] ".

This problem should be classified as: 1. Object Recognition.
The coordinates is [0.23, 0.34, 0.32, 0.46].

You are a helpful, highly intelligent guided assistant. You will do your best to help humans. Here are some examples:
... ...
You should learn how to better answer this question: ...

First ... Then... Finally...
So the answer is 25.

What time does the phone in the photo show?

There are now eight categories of problems: ......
Please determine which category the problem belongs to.
Afterwards, please analyze the provided image and focus on the specific area related to the question. Generate bounding box coordinates: [x1, y1, x2, y2] ".

This problem should be classified as: 7. Text Recognition.
The coordinates is [0.75, 0.13, 0.82, 0.18].

You are a helpful, highly intelligent guided assistant. You will do your best to help humans. Here are some examples:
... ...
You should learn how to better answer this question: ...

First ... Then... Finally...
So the answer is 1:45pm.

Figure 3: Case study when using DCoT.

LMMs and benchmark datasets. The results of the experiment and the ablation study are then discussed in detail. All experiments were conducted on a server with 8 NVIDIA RTX 3090 GPUs and performed using LLaVA-1.5 Liu et al. (2024a) and Qwen-VL-Chat Bai et al. (2023) and followed their default hyperparameters, unless otherwise stated. We conducted extensive experiments to evaluate the effectiveness of DCOT.

4.1. Datasets

ScienceQA Lu et al. (2022) is the first multimodal science question and answer dataset with 26 topics, 127 categories, and 379 skills, covering a wide range of domains, including multiple-choice questions in more than 20,000 science subjects.
VizWiz Gurari et al. (2018) collects datasets from real blind users, including more than 30,000 real visual questions with blurry noise data, each question gets 10 reference answers through crowdsourcing, which can accurately reflect the real needs of users, and can also reflect many trivial and small problems in the real implementation of VQA tasks.
TextVQA Singh et al. (2019) is used to benchmark visual reasoning based on text in images. More than 45,000 questions with 28,000 images require the model to read and reason about the text in the images in order to answer the questions.
MM-Bench Liu et al. (2023b) is curated multimodal dataset of approximately 3,000 questions across 20 capability dimensions enables the evaluation of model performance at a more granular level.
MM-Vet Yu et al. (2024) a total of 217 questions, it defines 6 core visual language functions: recognition, OCR, knowledge, language generation, spatial perception, and mathematical computing, and proposes an LLMs-based open output evaluator that can evaluate different question types and answer styles, resulting in a unified scoring metric.

4.2. Models

LLaVA-1.5 Liu et al. (2024a): Linear projection is replaced by an MLP projection layer that maps visual features into an embedded space shared with LLM. Using a visual encoder with CLIP-ViT-L 336 × 336 resolution, Vicuna was used as a language decoder, prompting the LMM's capabilities with visual instruction tuning. We conducted experiments using models with 7B and 13B parameters.

Qwen-VL-Chat Bai et al. (2023): The model is based on Qwen-7B, uses Vision Transformer as the visual encoder, and introduces a location-aware visual language adapter to compress the image feature sequence, maintaining fine-grained visual comprehension ability. Optimize your model's performance with a three-stage training process of pre-training, multi-task pre-training, and supervised fine-tuning.

4.3. Baselines

In our experiments, we compared our DCoT method with the other two advanced multimodal Chain-of-Thought method baselines as shown in Table 3, and to evaluate the added benefit of our approach to the pre-trained LMMs, our first baseline was to apply the model to the benchmark without any prompt engineering.

CCoT Mitra et al. (2024): It is a Zero-Shot Chain-of-Thought method that uses LMMs to generate a scene graph SG, describe the objects in the picture and their relationships and attributes, obtain structured data in Json format, and then use the scene graph data in json format in the prompt to get a response.

DDCoT Zheng et al. (2023): This prompt maintains a critical attitude with negative spatial prompts and first divides the reasoning responsibilities of the LLMs into reasoning and identification. The reasoning process takes the form of decomposing sub-problems, converting the form of reasoning into two-step reasoning, for sub problems, first provide reasons, and then use these reasons to infer the overall problem again.

4.4. Results

Table 1: The results of different retrieval strategies on ScienceQA and TextVQA.

Models	Method	Dataset	
		ScienceQA	TextVQA
LLaVA-1.5-7B	Image-Text	68.9	58.8
	Text-Only	**69.3**	**59.1**
	Image-Only	68.4	58.7
LLaVA-1.5-13B	Image-Text	74.1	62.0
	Text-Only	**74.8**	**62.4**
	Image-Only	73.5	61.6
Qwen-VL-Chat	Image-Text	71.1	62.0
	Text-Only	**71.5**	**62.1**
	Image-Only	70.6	61.7

Selection of search strategy. Since the effectiveness of In-Context Learning is highly dependent on the quality of prompts and the selection of examples. Therefore, in the retrieval phase of the demonstration example, we conducted a series of experiments with different retrieval strategies. These include: (i) Image-Text, which fuses image and text features for multimodal retrieval. (ii) Text-Only, which relies solely on text embedding for retrieval. (iii) Image-Only, only image features are used for retrieval.

The results, shown in Table 1, show that text-only retrieval improves the performance of scienceQA and TextVQA by 0.5 and 0.27, respectively, compared to image-text multimodal retrieval. It demonstrates that the traditional multimodal retrieval approach uses a text encoder and a visual encoder to encode the question and image to obtain a common embedding, this retrieval strategy is not a good choice for contextual examples when searching in very small data samples (our experiment only had 10 samples per category). when image-text multimodal retrieval is used, the correlation between the retrieved contextual examples and the actual query is weak, which undermines the final QA performance. This weakness is due to the large amount of information in the image that is not relevant to the question, which may dominate the embedding space and distract from the core query elements. In addition, as hypothesized by Winterbottom et al. (2020) for model biasing, where large multimodal models rely more on textual information when trained, or where the model itself is more mature and refined in its processing of text inputs, then using only text vectors may naturally lead to better performance. Therefore, our subsequent experiments use the (ii) Text-Only approach.

Table 2: Comparison of TopK Performance between LLaVA and Qwen VL Chat Models on ScienceQA and TextVQA Datasets.

Models	ScienceQA				TextVQA			
	K=1	K=2	K=3	K=4	K=1	K=2	K=3	K=4
LLaVA-1.5-7B	68.5	69.0	69.3	69.2	58.7	59.0	59.1	59.1
LLaVA-1.5-13B	73.7	74.4	74.8	74.5	61.8	62.2	62.4	62.4
Qwen-VL-Chat	70.8	71.3	71.5	71.5	61.4	61.9	62.1	62.1

The number of examples in context. As shown in Table 2, for ScienceQA we find that with the increase of K value, the accuracy first increases and then tends to stabilize or even decrease, which shows that increasing the number of K times will improve the performance of the model using context learning, but increasing the K value excessively can lead to performance degradation, which means that additional examples may introduce useless information, indicating a potential problem with redundancy in example selection. For TextVQA the performance will hardly change when the K value is continuously increased. This is due to the fact that the LLaVA-1.5 and Qwen-VL-Chat have a context window of only 4096, which makes it impossible to host too many demo examples. When K takes 3, the average improvement of the model in ScienceQA and TextVQA is 0.87 and 0.57, respectively, compared with only one example. Therefore, we used K=3 for all subsequent experiments.

Table 3: Main results table on MMBench, MM-Vet, TextVQA and VizWiz Benchmarks. Contains basic model scoring without any methodology and two advanced CoT methods as well as our methodology. Bold markings is the highest result, underline is the second highest result.

Method	MMBench	MM-Vet	ScienceQA	TextVQA	VizWiz
LLaVA-1.5-7B	64.3	31.1	66.8	58.2	50.0
LLaVA-1.5-7B+CCoT	**66.9**	32.0	68.7	58.9	51.1
LLaVA-1.5-7B+DDCoT	64.6	31.3	67.6	57.9	50.5
LLaVA-1.5-7B+DCoT (ours)	66.1	**32.9**	**69.3**	**59.1**	**51.9**
LLaVA-1.5-13B	67.7	36.1	71.6	61.3	53.6
LLaVA-1.5-13B+CCoT	**70.7**	37.5	73.9	61.9	54.3
LLaVA-1.5-13B+DDCoT	68.2	36.6	72.4	61.3	54.0
LLaVA-1.5-13B+DCoT (ours)	69.4	**38.4**	**74.8**	**62.4**	**55.7**
Qwen-VL-Chat	60.6	28.7	68.2	61.5	38.9
Qwen-VL-Chat+CCoT	63.9	29.6	71.0	61.7	41.1
Qwen-VL-Chat+DDCoT	61.9	29.5	70.5	61.1	40.6
Qwen-VL-Chat+DCoT (ours)	**64.5**	**30.1**	**71.5**	**62.1**	**42.7**

Main results. We demonstrate that after a fair comparison of the same demo examples, the experimental results of several benchmark datasets are shown in Table 3, Compared with the basic model, our DCoT method has an average improvement of 2.47, 1.83, 3, 0.87 and 2.6 respectively on MMBench, MM Vet, ScienceQA, TextVQA and VizWiz. Compared with CCoT's method, the average increase of MM Vet, ScienceQA, TextVQA and VizWiz is 0.77, 0.67, 0.37 and 1.27. Compared with DDCoT method, the average increase of MMBench, MM-Vet, ScienceQA, TextVQA and VizWiz was 1.77, 1.33, 1.87, 1.1 and 1.73 respectively, highlighting the effectiveness of our method and the scalability without additional computational overhead, And the comparative experiments of LLava-1.5-7B and LLaVA-1.5-13B show that the larger the number of parameters, the more obvious the improvement effect. These results show that the proposed DCoT can effectively improve the general performance of the model in different scenarios.

4.5. Ablations

The ablation study as shown in Table 4. Includes the two variations: a text guide that lacks fine-grained images, is based entirely on the original input images at the visual level, and implements contextual learning only by demonstrating examples; One that lacks contextual text guidance but inputs fine-grained images containing bounding boxes at the visual level for guidance.

On MMBench, MM-Vet and ScienceQA datasets, the average accuracy of ICL guidance was 0.83, 0.37 and 0.8 higher than that of FGI guidance, respectively, while on TextVQA and VizWiz datasets, the average accuracy of FGI guidance was 0.5 and 0.6 higher than that of ICL guidance, respectively. This shows that the two have different effects. If the answer exists explicitly within the image, without requiring any inference, the Fine-Grained Image

Table 4: The ablation study of the two main components of our proposed method: Fine-Grained Image and In-context learning where ✗ means we do not employ the corresponding method in our approach and ✓ means we use the corresponding method in the system. FGI: Fine-Grained Image, ICL: In-context learning.

Models	FGI	ICL	MMBench	MM-Vet	ScienceQA	TextVQA	VizWiz
Llava-1.5-7B	✓	✓	**66.1**	**32.9**	**69.3**	**59.1**	**51.9**
	✗	✓	65.1	32.0	69.1	58.4	50.3
	✓	✗	64.6	31.4	68.5	58.9	51.2
	✗	✗	64.3	31.1	66.8	58.2	50.0
Llava-1.5-13B	✓	✓	**69.4**	**38.4**	**74.8**	**62.4**	**55.7**
	✗	✓	68.3	37.2	73.6	61.6	54.7
	✓	✗	67.2	36.8	72.2	62.0	55.2
	✗	✗	67.7	36.1	71.6	61.3	53.6
Qwen-VL-Chat	✓	✓	**64.5**	**30.1**	**71.5**	**62.1**	**42.7**
	✗	✓	62.9	29.5	71.1	61.5	41.8
	✓	✗	62.0	29.4	70.7	62.0	42.2
	✗	✗	60.6	28.7	68.2	61.5	38.9

(FGI) module demonstrates better performance. Conversely, when the answer is implicit embedded within the image, necessitating logical reasoning, the In-context learning (ICL) module becomes essential to decipher the underlying logic. In terms of the overall results, the effect of the two modules together has the highest accuracy on the five data sets, and the average accuracy of the model reaches 66.7, 33.8, 71.9, 61.2 and 50.1, respectively.

5. Conclusion

In this paper, we propose an innovative Dual Chain-of-Thought (DCoT) strategy and a Fast In-Context Retrieval Framework (FICRF) to improve the inference performance of Large Multimodal Models in complex multimodal tasks. The DCoT strategy provides a plug-and-play performance enhancement scheme for LMMs through its unique two-pronged approach by combining fine-grained image-guided and contextual learning text-guided approach, through which the model can focus more precisely on the image area related to the problem and reduce the interference of irrelevant information. At the same time, the application of the Fast Unsupervised Retrieval Framework enables the text guidance stage to dynamically retrieve the most relevant examples according to the problem, which further enhances the logical reasoning ability of the model. Our work provides valuable insights and feasible solutions for the development of future multimodal AI systems, especially in the pursuit of more efficient and intelligent alternatives to model fine-tuning.

Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2023YFB3710901).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Yuhang Cao, Pan Zhang, Xiaoyi Dong, Dahua Lin, and Jiaqi Wang. Dualfocus: Integrating macro and micro perspectives in multi-modal large language models. *arXiv preprint arXiv:2402.14767*, 2024.

Julian Coda-Forno, Marcel Binz, Zeynep Akata, Matt Botvinick, Jane Wang, and Eric Schulz. Meta-in-context learning in large language models. *Advances in Neural Information Processing Systems*, 36:65189–65201, 2023.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

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.

Jannik Kossen, Yarin Gal, and Tom Rainforth. In-context learning learns label relationships but is not conventional learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *Association for Computational Linguistics*, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing:292–305, 2023b.

Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, and Zhongyu Wei. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. *arXiv preprint arXiv:2405.16919*, 2024a.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773, 2024b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Weihao Liu, Fangyu Lei, Tongxu Luo, Jiahe Lei, Shizhu He, Jun Zhao, and Kang Liu. Mmhqa-icl: Multimodal in-context learning for hybrid question answering over text, tables and images. *arXiv preprint arXiv:2309.04790*, 2023a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023b.

Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*, 2024c.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Yang Luo, Zangwei Zheng, Zirui Zhu, and Yang You. How does the textual information affect the retrieval of multimodal in-context learning? *arXiv preprint arXiv:2404.12866*, 2024.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024.

Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14076–14088, 2024.

Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, pages 1046–1056. PMLR, 2022.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.

David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. *arXiv preprint arXiv:2403.02325*, 2024.

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.

Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. On modality bias in the tvqa dataset. *arXiv preprint arXiv:2012.10210*, 2020.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.

Yifan Wu, Pengchuan Zhang, Wenhan Xiong, Barlas Oguz, James C Gee, and Yixin Nie. The role of chain-of-thought in complex vision-language reasoning task. *arXiv preprint arXiv:2311.09193*, 2023.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International conference on machine learning*. PMLR, 2024.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*, pages 202–227. PMLR, 2024.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36, 2024.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.