Interleaved-Modal Chain-of-Thought

Jun Gao[1],Yongqi Li[2]*, Ziqiang Cao[1]*, Wenjie Li[2]

[1]School of Computer Science and Technology, Soochow University
[2]Department of Computer Science, The Hong Kong Polytechnic University
jgao1106@stu.suda.edu.cn, liyongqi0@gmail.com
zqcao@suda.edu.cn, cswjli@comp.polyu.edu.hk
https://github.com/jungao1106/ICoT

Abstract

Chain-of-Thought (CoT) prompting elicits large language models (LLMs) to produce a series of intermediate reasoning steps before arriving at the final answer. However, when transitioning to vision-language models (VLMs), their text-only rationales struggle to express the fine-grained associations with the original image. In this paper, we propose an image-incorporated multimodal Chain-of-Thought, named ***Interleaved-modal Chain-of-Thought (ICoT)****, which generates sequential reasoning steps consisting of paired visual and textual rationales to infer the final answer. Intuitively, the novel ICoT requires VLMs to enable the generation of fine-grained interleaved-modal content, which is hard for current VLMs to fulfill. Considering that the required visual information is usually part of the input image, we propose* ***Attention-driven Selection (ADS)*** *to realize ICoT over existing VLMs. ADS intelligently inserts regions of the input image to generate the interleaved-modal reasoning steps with ignorable additional latency. ADS relies solely on the attention map of VLMs without the need for parameterization, and therefore it is a plug-and-play strategy that can be generalized to a spectrum of VLMs. We apply ADS to realize ICoT on two popular VLMs of different architectures. Extensive evaluations of three benchmarks have shown that ICoT prompting achieves substantial performance (up to 14%) and interpretability improvements compared to existing multimodal CoT prompting methods.*

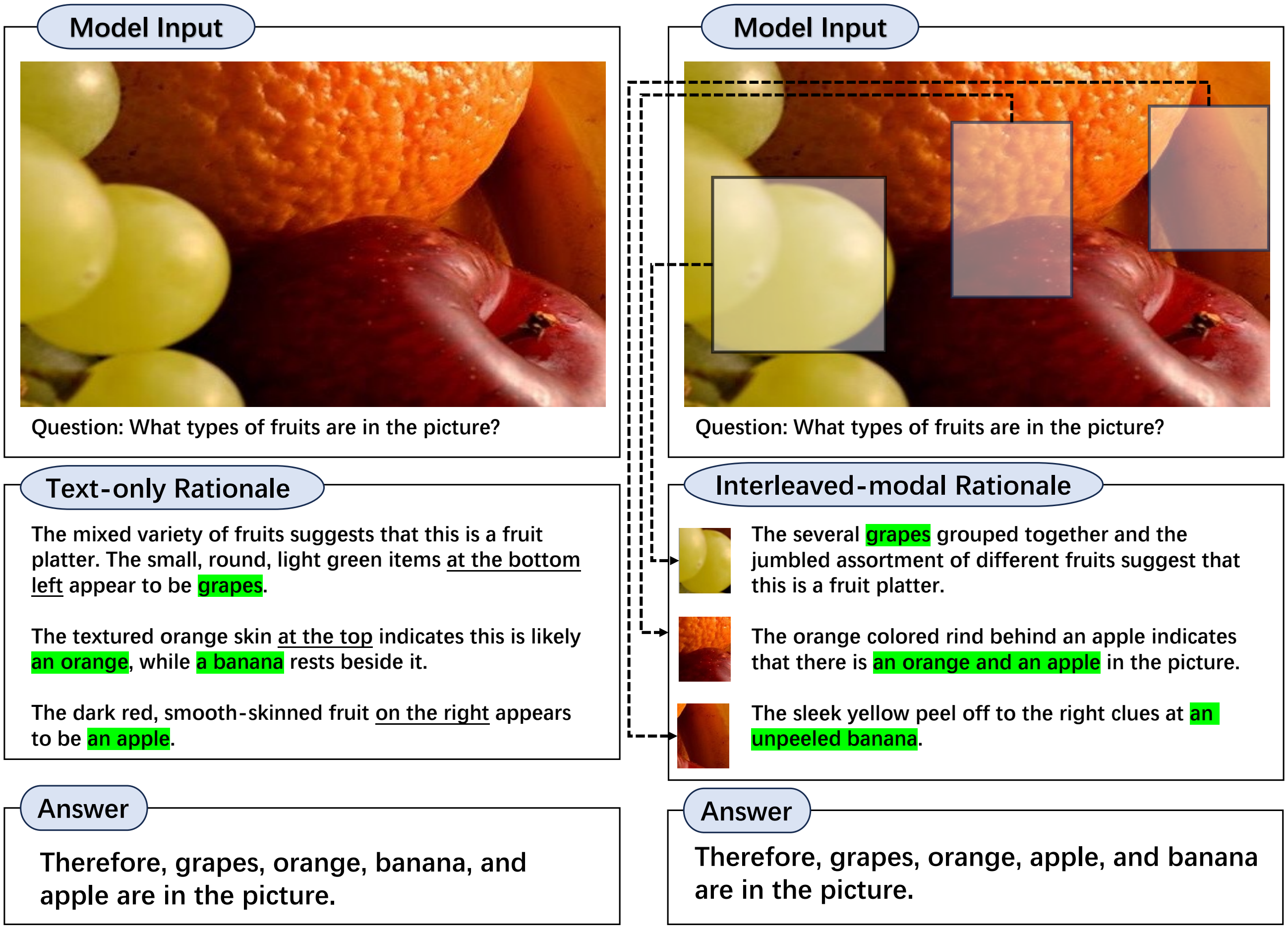
1. Introduction

Chain-of-Thought (CoT) [31] prompting aims to augment the reasoning capabilities of large language models (LLMs) [4, 8, 22, 26] by eliciting them to produce a sequence of intermediate natural language reasoning steps before arriving at the final output. CoT has proven effective in various reasoning tasks, including arithmetic [9], commonsense [17], and symbolic [3], and it has become a potential pathway to advanced artificial intelligence, as depicted in GPT-o1 [20].

With the development of vision-language models (VLMs), extending CoT prompting into multimodal CoT to improve the reasoning capabilities of VLMs in vision-related tasks becomes increasingly important [19, 28, 33, 34]. The initial multimodal CoT attempts [28, 33] take as input the fused visual and textual embeddings, and train language models, such as T5 [23] models, to generate text-only rationales and answers. In the era of VLMs, introducing triple demonstrations composed of an image with the instruction, textual rationales, and the final output (e.g., answer), has proven effective in sparking the reasoning ability of VLMs [6]. Then, related studies focus on improving the linguistic reasoning ability of VLMs. Specifically, DDCoT [34] leverages VLMs to deconstruct problems and resolve them respectively, and CCoT [19] generates scene graphs to prompt VLMs with object and position description. SCAFFOLD [12] overlays a coordinate matrix onto the image to prompt the VLMs with relative visual positions. However, these methods still generate text-only reasoning steps, making it hard to express the fine-grained associations with the origin image exactly. As shown on the left of Figure 1, textual position descriptions, e.g., at the top, are too rough to identify all fruits (orange and banana).

In light of the limitations of text-only rationales, we propose incorporating visual information to enhance the precision of fine-grained associations between generated textual rationales and the corresponding image. We therefore propose an advanced multimodal Chain-of-Thought prompting, named **Interleaved-modal Chain-of-thought (ICoT)**, as shown on the right of Figure 1. ICoT generates multimodal rationales consisting of paired images and textual rationales that formulate interleaved-modal reasoning steps

*Corresponding authors.



Question: What types of fruits are in the picture?

The mixed variety of fruits suggests that this is a fruit platter. The small, round, light green items at the bottom left appear to be grapes.

The textured orange skin at the top indicates this is likely an orange, while a banana rests beside it.

The dark red, smooth-skinned fruit on the right appears to be an apple.

Therefore, grapes, orange, banana, and apple are in the picture.

Question: What types of fruits are in the picture?

The several grapes grouped together and the jumbled assortment of different fruits suggest that this is a fruit platter.

The orange colored rind behind an apple indicates that there is an orange and an apple in the picture.

The sleek yellow peel off to the right clues at an unpeeled banana.

Therefore, grapes, orange, apple, and banana are in the picture.

Figure 1. The illustration between multimodal CoT with text-only rationales (Left) and interleaved-modal rationales (Right). Green blocks are correct texts used to infer the final answer. Text-only rationales restrict VLMs to use a rough description to indicate the position of objects. Transparent boxes indicate that these regions are selected and inserted to formulate paired visual and textual rationales in ICoT.

to infer the final output. To the best of our knowledge, ICoT is the first multimodal CoT with images incorporated, and it aligns more closely with human thinking processes [5, 21].

Intuitively, facilitating the novel ICoT is non-trivial, as it introduces challenges for VLMs to support fine-grained interleaved-modal content generation. No current VLMs meet this condition completely. Perceiver-based VLMs such as Qwen2-VL [29] converts images into visual embeddings. Thus, they support fine-grained visual understanding but cannot generate multimodal outputs. Recently proposed unified-modeling VLMs, such as Chameleon [25], Unified-IO 2 [18], and Emu-3 [30], enable multimodal generation by tokenizing images into discrete tokens. However, on the one hand, unified-modeling VLMs exhibit inertia toward multimodal content generation [7]; on the other hand, the generated images belong to the fixed pre-defined resolution instead of fine granularity.

Since required visual information is usually part of the input image for ICoT, we accordingly propose **Attention-driven Selection (ADS)** to realize ICoT. The basic idea of ADS is to signal VLMs to select patches from the input image rather than generating extra images. At the beginning of generating each textual rationale, ADS inserts a piece of visual tokens of selected patches from the input image to refine the generation of the following textual rationale. Specifically, ADS utilizes the attention map of VLMs to identify optimal patches from the input image as fine-grained visual rationales. Once these fine-grained visual rationales are inserted into the current generation sequence, the VLM resumes the original autoregressive text generation process based on previous multimodal content, formulating paired image and textual rationales to infer the final outputs. Notably, since ADS does not compel VLMs to generate real images, it brings ignorable inference latency compared with previous text-only CoT methods. Additionally, ADS leverages the attention map of VLMs without requiring parameterization, making it a plug-and-play strategy that can be easily adapted to a wide range of VLMs.

In this paper, we apply **ADS** to realize **ICoT** on Chameleon and Qwen2-VL, representing the state-of-the-art unified modeling and perceiver-based VLMs. The results on existing datasets, including M^{3}CoT [6], ScienceQA [24], and LLaVA-W [15], indicate that ICoT realized by ADS brings VLMs with substantial performance gains (up to 14%) compared with current multimodal CoT methods. Additionally, it is noted that the tracked interleaved-modal rationales further enhance the interpretability of the generated results. Our main contribution can be concluded as follows:

- We propose interleaved-modal CoT, which innovates text-only rationales into multimodal ones to construct clearer reasoning. To our knowledge, we are the first to incorporate images into the intermediate reasoning steps in multimodal CoT.
- We propose an effective and efficient Attention-driven Selection strategy to facilitate ICoT, which is training-free and widely applicable to VLMs without requiring them to support multimodal generation.
- Experiments demonstrate that our ICoT significantly surpasses existing multimodal CoT methods, proving that the interleaved-modal reasoning process is a foundational innovation in the line of CoT.

2. Related Work

2.1. Vision-Language Models (VLMs)

Currently, predominate VLMs such as Qwen-VL [2, 29], BLIP [13], and LLaVA [14–16] are mainly built upon a Large Language Model (LLM), a visual module, and an aligned vision-language adapter. The visual module, e.g., Vision Transformer (ViT) [1], encodes images into dense representations, and then the adapter, e.g., MLP or Q-Former, converts these representations into LLM-readable visual tokens. Finally, visual tokens and textual tokens are fed into the LLM to perform the next-token prediction. This type of VLM can be concluded as Perceiver-LLM architecture, while the Perceiver usually comprises a visual module and the adapter. Additionally, Cambrain-1 [27] introduces more visual modules to collaboratively provide more useful visual tokens in a vision-centric paradigm. In the other research line, unified-modeling VLMs represented by Chameleon [25], Unified-IO 2 [18], and Emu3 [30] are designed to generate texts, images, and so on uniformly. As these models apply codebook [11] to tokenize images into discrete vokens, their training processes are supervised by both vision and text information. Unified-modeling VLMs are expected to develop more stable multimodal understanding abality [10].

2.2. Multimodal Chain-of-Thought Prompting

Similar to CoT used in LLMs, multimodal Chain-of-Thought prompting methods [12, 19, 28, 32, 34] aim to augment the reasoning ability of VLM by generating intermediate reasoning steps. A series of studies focus on providing VLMs with fine-grained textual information, such as detailed description [28]. Compositional CoT (CCoT) [19] prompts VLMs to generate a Scene Graph (SG), which is a JSON-like description containing compositional information of objects that occurred in the image. DDCoT [34] deconstruct problems into small problems, requiring VLMs to solve them respectively and then inferring the final answer. In the other research line, Set-of-Marks prompting [32] augments the objects in the image to help VLMs recognize them. SCAFFOLD [12] overlays coordinate onto images to prompt VLMs with relative position information, and VLMs leverage overlayed textual coordinates to implicitly represent corresponding regions of the image to perform reasoning.

However, these methods still produce text-only rationales to infer the final answer. These generated textual rationales usually struggle to express the fine-grained associations with the origin image. We thereby propose ICoT to elicit VLMs to generate interleaved visual-textual reasoning steps to effectively reach the final outputs.

3. Methodology

To address the limitations that current multimodal CoT methods are still stuck in generating text-only rationales to infer the final answer, we propose interleaved-modal CoT (ICoT) to elicit VLMs generated multimodal reasoning steps. We start by introducing the workflow of VLMs and multimodal CoT in Section 3.1. We then introduce the concept of ICoT in Section 3.2. Finally, we propose a plug-and-play method, Attention-driven Selection (ADS), to realize ICoT on existing VLMs.

3.1. Preliminaries

We first recall some background of VLMs and multimodal CoT in this section.

Vision-Language Model. VLMs usually consist of a visual encoder $\mathbf{E}$ and a generative large language model LLM, and they determine where to insert images according to visual holders inserted in the text instructions. Then, VLMs take the image and the instructions as input and respond with a final answer

$$\text{answer} = \mathbf{VLM}(\text{Image}, \text{Instruction}). \tag{1}$$

Specifically, the visual encoder $\mathbf{E}$ extracts visual tokens $f_v^{l\times d}$ from the image x_v, where l is the length of visual

Algorithm 1 Interleaved-modal CoT

```
Input: Word embeddings $f_e$, Visual tokens $f_v$, Selected number $n$, Signal tokens $\mathcal{S}$, Stopping criteria $SC$
Output: Generated Response $Answer$
predicted_tokens ← []    ▷ Initialize as an empty list
$\mathcal{V} \leftarrow f_v$
inputs = Initilize($f_e$, $f_v$)    ▷ Initialize inputs for prefilling
while $SC$ not met do
    next_token, attention_map = model(inputs )
    Append next_token to predicted_tokens
                                    ▷ ADS judgement
    if predicted_tokens $= \mathcal{S}$ then
        $\mathcal{V}_{\text{selected}}$= ADS($\mathcal{V}$, attention_map, $n$)    ▷ Apply Attention-driven Selection
        Append $\mathcal{V}_{\text{selected}}$ to predicted_tokens
    end if
    inputs = Update(inputs, predicted_tokens)    ▷ Updates inputs for next step generation
end while
$Answer$ = Tokenizer.decode(predicted_tokens)
return $Answer$
```

Algorithm 2 Attention-driven Selection

```
Input: Attention map $A_t$, Selected number $n$, Visual tokens $f_v$
Output: Fine-grained visual information $\mathcal{V}_{\text{selected}}$
$\mathcal{V}_{\text{selected}} \leftarrow \emptyset$    ▷ Initialize as an empty set
Indices ← $\text{TopK}(A_t, n)$
for $i$ in Indices do
    Append $f_v^{l \times d}[i]$ to $\mathcal{V}_{\text{selected}}$
end for
Restore($\mathcal{V}_{\text{selected}}$, Indices)    ▷ Restore relative positions in-place
return $\mathcal{V}_{\text{selected}}$
```

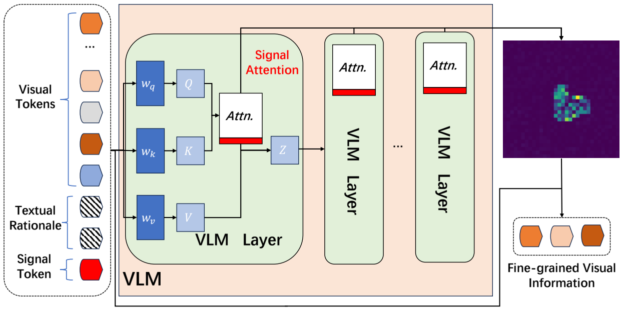


Figure 2. The workflow of ADS selecting fine-grained visual rationales. Signal attention represents the attention map of the signal token overall visual tokens.

tokens and d is the dimensions of the hidden states of the LLM. The built-in LLM predicts next-tokens in a left-to-right fashion according to visual tokens $f_v^{l \times d}$ and the instructions.

Multimodal CoT. Compared with the direct prediction described in Eqn. 1, Multimodal CoT further introduces a prompt to elicit VLMs to generate a series of intermediate textual rationales $(r_1, r_2, ...)$ before the final answer:

$$r_1, r_2, ..., \text{answer} = \textbf{VLM}(\text{Prompt}, \text{Image}, \text{Instruction}). \quad (2)$$

Technically, the prompt could be represented as a sequence of demonstrations, each consisting of a triple: $(\text{Image}, \text{Rationale}, \text{Answer})$. Alternatively, an explicit instruction, such as “Let’s think step by step,” could also serve as the prompt.

3.2. Interleaved-modal Chain-of-Thought

Previous multimodal CoT prompting methods only produce text-only reasoning steps to improve the reasoning ability of VLMs. These intermediate steps are generated according to the entire image, which are struggle to express exact fine-grained associations with the original image. Given these limitations, we propose a more advanced Interleaved-modal Chain-of-Thought (ICoT) prompting, aiming to elicit VLMs to generate a series of multimodal intermediate reasoning steps each consisting of paired image and textual rationale. Generated intermediate reasoning steps formulating interleaved-modal rationales to effectively lead to the final outputs. In this paper, we consider the visual rationales in interleaved-modal rationales as fine-grained visual information $x_v^{h' \times w'}$ extracted from an image $x_v^{h \times w}$[1]. These visual rationales capture relevant details in the image, such as objects, colors, and texts, interleaved with the following generated textual rationale to infer the final answer:

$$r_1, x_{v_1}, r_2, x_{v_2}, ..., \text{answer} = \textbf{VLM}(\text{Prompt}, \text{Image}, \text{Instruction}). \quad (3)$$

3.3. Attention-driven Selection

Although the proposed ICoT is both novel and conceptually sound, current VLMs are unable to generate such fine-grained visual information. This remains true even for VLMs [18, 25, 30] that are empowered with multimodal generation ability. We thus propose to simplify the problem from fine-grained visual information generation to fine-grained visual information selection, as this information

[1] The images in the dataset involved in this paper are RGB images by default, and the number of channels is omitted in the formulas for simplicity.

has been naturally contained in the origin image, namely $x_v^{h' \times w'} \in x_v^{h \times w}$ where $h' \ll h$ and $w' \ll w$.

Specifically, before performing next-token prediction, ICoT requires the VLM to cache visual tokens $f_v^{l \times d}$ extracted by its built-in visual encoder $\mathbf{E}(x_v)$ for further selection. In the following decoding steps, we consider the VLM deems it necessary to insert a piece of visual rationales after generating a pre-defined signal token $\mathcal{S}$ as shown in Figure 2, which is a natural language token that indicates the beginning of a textual rationale. Therefore, ADS will be signaled to select fine-grained visual information from $f_v^{l \times d}$ upon the VLM generating $\mathcal{S}$:

$$\text{do\_selection} = \begin{cases} \text{True} & \text{if predicted\_tokens}[-1] = \mathcal{S}; \\ \text{False} & \text{otherwise.} \end{cases} \tag{4}$$

Then, ADS selects n visual tokens from $f_v^{l \times d}$ as fine-grained visual information according to the signal attention map A_t at the current decoding step t:

$$\mathcal{V}_{\text{selected}} = \{f_v[i] | i \in \text{TopK}(A_t, n)\}, \tag{5}$$

where A_t is obtained by averaging the attention map between the signal token and visual tokens across all VLM layers. Up to now, the selected visual tokens are sorted by their attention scores, and we subsequently restore the relative position of $f_v[i]$ in the origin image in place, prioritizing rows. Once fine-grained visual tokens $\mathcal{V}_{\text{selected}}$ are obtained, VLMs will take as input the concatednated $\mathcal{V}_{\text{selected}}$ and predicted_tokens, denote as $\text{Cat}(\text{predicted\_tokens}, \mathcal{V}_{\text{selected}})$, resuming the original autoregressive text generation process. Hence, the current decoding step is formulated as:

$$\begin{aligned} \text{next\_token} = \mathbf{VLM}(\text{Cat}(&\text{predicted\_tokens}, \\ &\mathcal{V}_{\text{selected}})). \end{aligned} \tag{6}$$

Notably, to avoid misunderstanding, we first convert predicted_tokens into word embeddings f_e, and then concatenate f_e with selected fine-grained visual information $\mathcal{V}_{\text{selected}}$ in the embedding-level. We provide a detailed description of ICoT and ADS in Algorithm 1 and Algorithm 2, respectively.

Technically, ICoT inherits the existing eliciting methods from CoT, such as attaching an instruction: “Let’s think step by step” in zero-shot ICoT or providing few-shot examples. In the few-shot ICoT, each example consists of an image, textual rationales, and visual rationales. These examples can be manually designed to prompt VLMs regarding their way of thinking and generation formatting, among other aspects.

4. Experiments

4.1. Datasets

M^{3}CoT [6] is a novel multimodal CoT benchmark specifically concentrated on multi-domain, multi-reasoning-step. M^{3}CoT contains 267 categories from science, mathematics, and commonsense domains. As the question of each instance is relatively complex, their rationales have an average length of 293 tokens and rely more on fine-grained visual information, which can reflect the advantages of ICoT compared with previous multimodal CoT methods.

ScienceQA [24] is a popular dataset used to evaluate the reasoning ability of VLMs. We use ScienceQA to provide a general comparison between ICoT and other existing multimodal CoT methods.

LLaVA-Bench In-the-Wild (LLaVA-W) [15] evaluates VLMs’ ability to respond to visual questions with detailed long-form answers, which also focus on the fine-grained visual description. The reference label of each instance is produced by GPT-4v.

4.2. Baselines

No-CoT responds to the current input image and question directly without further prompting. The few-shot demonstrations of the direct generation mode consist of (Image, Question, Answer).

Multimodal CoT [33] elicits VLMs to generate a series of text-only intermediate reasoning steps to infer the final outputs.

CCoT [19] first generates a scene graph (SG) using the VLM itself and then uses that SG in the prompt to produce a response. The SG is a JSON-like structural description of the given image with extensive compositional information of objects in the current images. Following their settings, we apply their official prompt to prompt VLMs to generate SGs and answers respectively.

DDCoT [34] first prompts LLM to deconstruct the input question into a sequence of basic sub-questions and then applies a VQA model to answer these sub-questions involving visual information. In this paper, we use the VLM that plays the role of LLM in DDCoT for a fair comparison, as their original LLM is ChatGPT.

SCAFFOLD [12] overlays a coordinate matrix onto the input image, exactly demonstrating relative visual positions for VLMs. During reasoning, VLMs are steered to utilize these coordinates that indicate fine-grained visual information in the image to solve different vision-language tasks. We use their released scripts to add coordinates over each image and then use their official prompt to elicit VLMs.

Specifically in the few-shot scenario, the demonstrations of these baselines are human-written, aligning with ICoT.

Backbone	Methods	*0-shot*			*1-shot*		
		M^3CoT ACC. ↑	ScienceQA ACC. ↑	LLaVA-W ROUGE-L↑	M^3CoT ACC. ↑	ScienceQA ACC. ↑	LLaVA-W ROUGE-L↑
Chameleon-7B	No-CoT	29.1	47.7	13.1	28.4	48.5	23.9
	Multimodal CoT [33]	28.5	49.0	20.4	30.6	50.7	20.6
	CCoT [19]	29.4	50.2	22.1	31.4	51.3	24.5
	DDCoT [34]	28.6	49.8	20.2	29.8	49.2	23.1
	SCAFFOLD [12]	29.6	48.5	21.7	31.1	47.5	24.7
	ICoT (Ours)	**29.8**	**51.0**	**25.2**	**32.3**	**53.4**	**27.6**
	% **Improve**	0.6%	1.6%	14.0%	2.8%	4.0 %	11.7%

Backbone	Methods	*0-shot*			*1-shot*		
		M^3CoT ACC. ↑	ScienceQA ACC. ↑	LLaVA-W ROUGE-L↑	M^3CoT ACC. ↑	ScienceQA ACC. ↑	LLaVA-W ROUGE-L↑
Qwen2-VL-7B	No-CoT	43.6	56.3	32.7	45.4	64.4	33.5
	Multimodal CoT [33]	40.1	51.3	30.7	42.5	58.3	31.4
	CCoT [19]	43.3	56.4	29.4	44.1	63.8	33.9
	DDCoT [34]	42.6	55.2	31.2	45.7	64.9	32.8
	SCAFFOLD [12]	41.7	53.7	31.8	44.9	62.5	33.1
	ICoT (Ours)	**44.1**	**56.8**	**34.2**	**46.0**	**65.4**	**35.7**
	% **Improve**	1.1%	0.7%	4.6%	0.6%	0.7%	5.3%

Table 1. Results of ICoT and baselines based on Chameleon and Qwen2-VL, with the highest score **bold**. M^3CoT and ScienceQA are evaluated by accuracy, and we report the ROUGE-L score for the LLaVA-W benchmark. % **improve** represents the relative improvement achieved by ICoT over the previously best baseline.

4.3. Implement Details

We apply ICoT over Chameleon-7B [25] and Qwen2-VL-7B-Instruct [29], which represents the fine-grained visual information in the form of discrete vokens and dense features. All experiments are conducted on A800 GPUs, and we evaluate ICoT under both zero- and one-shot scenarios. During generating interleaved-modal rationales, the signal token $\mathcal{S}$ used to trigger ADS is set to line break, i.e., "\n", by default, which semantically and empirically indicates the end of a generated rationale and the beginning of the next one.

VLMs insert visual tokens selected by ADS following the special token at the granularity of 64 according to posterior results shown in Table 4 of Appendix 8. Notably, to shorten the representation of an image, Qwen2-VL introduces a novel merge mechanism, and we approximately consider its patch size to be (28×28). Each patch of Qwen2-VL has approximately 4 times as many pixels as a Chameleon patch (16×16), which results in practical selection numbers of ADS set to 16. Considering the work of ADS requires the inner attention map, we apply the "eager" attention on both Chameleon-7B and Qwen2-VL, limited to the dependency of related python libraries [2].

4.4. Main Results

We comprehensively evaluate the performance of ICoT on top of Chameleon-7B and Qwen2-VL-7B through M^3CoT, ScinceQA, and LLaVA-W in Table 1. In 0-shot settings, ICoT outperforms all baselines, including direct generation (No-CoT), CoT, CCoT, DDoT, and SCAFFOLD. Specifically, ICoT distinguishes from Multimodal CoT in terms of the modality of reasoning steps, which exhibit the advantages of interleaved-modal rationales to infer the final answer effectively. Compared with other multimodal CoT methods, the performance gains of ICoT further indicate that interleaved-modal rationales are more reasonable in intuition and effect than plainly inserted scene graphs (CCoT) and deconstructed sub-questions (DDoT). In 1-shot settings, ICoT demonstrations contain manually selected fine-grained visual information, while their text rationales are kept the same as other baselines. The performance gains compared with 0-shot ICoT indicate that our manually designed fine-grained ICoT demonstrations potentially guide VLMs to think in this format. In Table 4, we rigorously ablate the effectiveness of fine-grained ICoT demonstrations.

Additionally, ICoT achieves the most relative performance gains in the LLaVA-W benchmark as the reference labels contain details sourced from images. These substantial performance gains compared with other baselines prove that visual tokens selected by ADS effectively capture the fine-grained visual information of an image, aiding the gen-

[2] Using attn_implementation="eager" when loading the model from HuggingFace.

eration of high-quality text rationales.

4.5. Ablation Study

We ablate ICoT to verify the effectiveness of each portion across three benchmarks in Table 2 with the following settings: (1). w/o ADS: VLMs generate text-only rationales. (2). w/o FVI: Patches inserted in the demonstration are randomly sampled. Results indicate that both ADS and fine-grained visual information (FVI) incorporated in the demonstration are necessary. In particular, interleaved-modal rationales exhibit substantial advantages in generating high-quality textual rationales compared with text-only rationales (w/o ADS). When substitute ICoT demonstration with normal ones (w/o FVI), the performance degradation proves the fact that fine-grained visual information in demonstrations effectively guides VLMs to think in this format. Compared with the performance difference between removing ADS and FVI, we find that generating paired visual and text rationales boosts more improvements.

Additionally, the performance gap is relatively smooth in ScienceQA and more dramatic on M^{3}CoT and LLaVA-W. We attribute this to the ScienceQA dataset being relatively easier than others since both M^{3}CoT and the answers of LLaVA-W highly rely on the fine-grained visual information of an image. Therefore, our proposed ICoT has the potential to solve complex vision-language tasks.

4.6. In-depth Analysis

Analysis on realizing ICoT via KV Cache Up to now, the fine-grained visual information of ICoT is provided at the input end via discrete vokens or dense visual tokens, which brings more computation. After rethinking the generating process of an autoregressive model, there are other inputs that are proposed to avoid repeated computation, namely, the Key-Value (KV) Cache. The input image was stored in the KV Cache during the prefilling phase before generating multimodal intermediate reasoning steps in a left-to-right fashion. Therefore, copying the KV cache of fine-grained visual information enables ICoT with reduced computational costs, as visual information does not require extra forward propagation. As shown in Table 3, copying the KV Cache brings performance degradation compared with providing visual information at the input end. We attribute this phenomenon to the fact that although copying KV Cache indeed makes VLMs attend more to the same region as ADS, the optimal visual information is highlighted in a **position-agnostic** case, determined by the nature of KV Cache. Specifically, this degrades the original interleaved-modal rationales into non-interleaved ones as position information is early fused into KV Cache, and thus the copied ones are inherently insensitive to the position of textual rationale.

Methods	M^{3}CoT	ScienceQA	LLaVA-W
ICoT	32.3	53.4	27.6
w/o ADS	29.2 (-3.1)	52.4(-1.0)	24.5(-3.1)
w/o FVI	30.6 (-1.8)	52.8(-0.6)	25.9(-1.7)
w/o ADS+FVI	29.1(-3.2)	51.0(-2.4)	23.0(-4.6)

Table 2. Ablation studies of *1-shot* ICoT on Chameleon-7B. () describes the performance degradation compared with ICoT. FVI indicates the 1-shot demonstration contains fine-grained visual information. ADS indicates that VLMs generate interleaved-modal reasoning steps.

	0-shot		*1-shot*	
Dataset	KV-Copy	ICoT	KV-Copy	ICoT
M^{3}CoT	29.1	29.8	31.5	32.3
ScienceQA	49.7	51.0	52.9	53.4
LLaVA-W	24.7	25.2	27.0	27.6

Table 3. Results comparison between copying KV Cache (KV-Copy) and inserting selected patches.

Methods	M^{3}CoT	ScienceQA	LLaVA-W
Human-written	32.3	53.4	27.6
Model-written	31.5(-0.8)	51.8(-1.6)	26.7(-0.8)

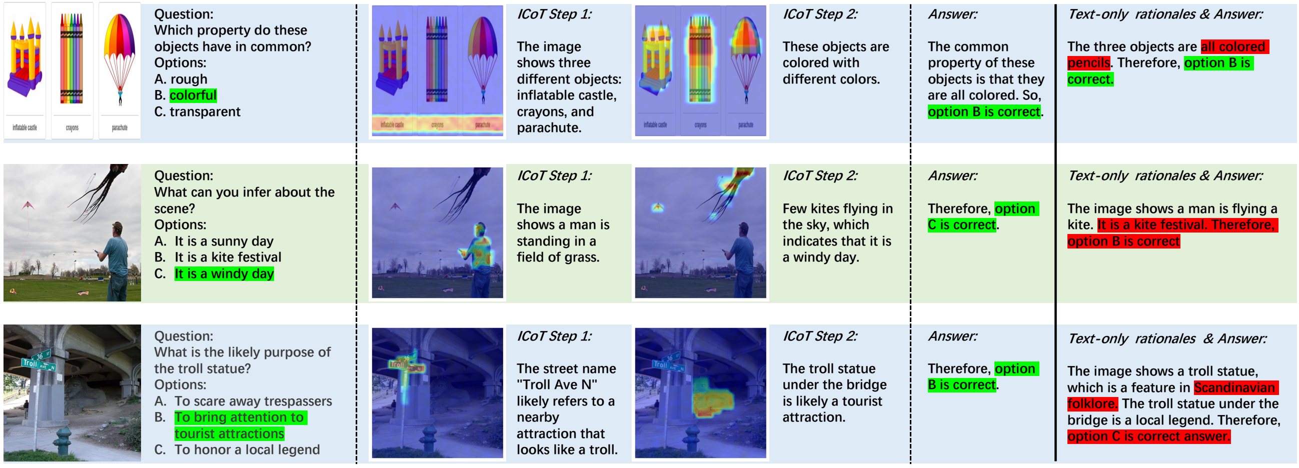
Table 4. Results concerning the design of demonstrations on Chameleon-7B. () describes the performance degradation compared with ICoT. Human-written indicates that demonstrations are manually designed with fine-grained visual information inserted, and Model-written indicates that the VLM generates both visual and textual rationale via ICoT.

However, considering it brought slight performance degradation and factually reduced computation costs, we believe this exploration is still valuable, and we call for more interesting exploration in realizing ICoT.

Analysis on the Demonstrations We also attempt to let VLMs generate the demonstrations via themselves (Automatic at the bottom of Table 4). Results indicate that using the automatically generated demonstrations also brings performance degradation compared with ICoT using manually designed ones. We consider it is caused by the fact that formulating a continuous sub-image through ADS is non-trivial, and some discrete patches inevitably introduce additional noise. Therefore, considering that designing such a demonstration is not time-consuming, ICoT utilizes manually designed ones to elicit VLMs to perform ICoT for better performance.

5. Case Study

In this section, we empirically illustrate the advantages of ICoT via three case studies in Figure 3. These case studies



Question:
Which property do these objects have in common?
Options:
A. rough
B. colorful
C. transparent

ICoT Step 1:
The image shows three different objects: inflatable castle, crayons, and parachute.

ICoT Step 2:
These objects are colored with different colors.

Answer:
The common property of these objects is that they are all colored. So, option B is correct.

Text-only rationales & Answer:
The three objects are all colored pencils. Therefore, option B is correct.

Question:
What can you infer about the scene?
Options:
A. It is a sunny day
B. It is a kite festival
C. It is a windy day

ICoT Step 1:
The image shows a man is standing in a field of grass.

ICoT Step 2:
Few kites flying in the sky, which indicates that it is a windy day.

Answer:
Therefore, option C is correct.

Text-only rationales & Answer:
The image shows a man is flying a kite. It is a kite festival. Therefore, option B is correct

Question:
What is the likely purpose of the troll statue?
Options:
A. To scare away trespassers
B. To bring attention to tourist attractions
C. To honor a local legend

ICoT Step 1:
The street name "Troll Ave N" likely refers to a nearby attraction that looks like a troll.

ICoT Step 2:
The troll statue under the bridge is likely a tourist attraction.

Answer:
Therefore, option B is correct.

Text-only rationales & Answer:
The image shows a troll statue, which is a feature in Scandinavian folklore. The troll statue under the bridge is a local legend. Therefore, option C is correct answer.

Figure 3. Case studies between ICoT and multimodal CoT with text-only rationales on Chameleon. Three cases are selected according to three typical problems in text-only problems: misunderstanding, overgeneralization, and hallucination. Red blocks indicates the incorrect rationales.

focused on three typical problems that occurred in text-only rationales, namely, misunderstanding (top), overgeneralization (middle), and hallucination (bottom).

Specifically, in the first case, interleaved-modal CoT first recognizes three different objects via captions: “inflatable castle, crayons, and a parachute”. Then, in the second reasoning step, ADS inserts selected patches from the scheduled objects to elicit the VLM to conclude their common property, and VLM infers a correct answer. Text-only CoT misunderstands the three objects are all colored pencils, ignoring the castle and the parachute, even the final answer is correct. In the second case, text-only rationales overgeneralize flying a kite to a kite festival, leading to a wrong answer. ICoT first recognizes a man standing in a field of grass and then infers it is a windy day according to a few kites in the sky. In the last case, it provides the other typical error of text-only CoT, namely, hallucination. As text-only CoT purely relies on language reasoning ability, VLMs have the potential to imagine something not mentioned in the image, resulting in a wrong answer. ICoT first infers from the street sign that there may be a troll attraction nearby according to patches of the indicator inserted by ADS. Then, the ADS helps the VLMs to attend to the troll statue under the bridge and infer it is likely the mentioned attraction, finally arriving at the correct answer.

Even though the above case studies exhibit the advantages of ICoT, ADS still brings potential problems. For example, ADS is triggered to select patches when VLM generates a pre-defined signal token. This simple mechanism is a double-edged sword that VLMs will generate low-quality responses if this token is generated with a high frequency.

6. Conclusion

In this paper, we first propose interleaved-modal CoT (ICoT), which generates interleaved-modal rationales to infer the final answer effectively. In light of the challenges of applying ICoT on existing VLMs, we then introduce Attention-driven Selection (ADS), a plug-and-play strategy to identify optimal patches from the image without being parameterized. We evaluate ICoT on Chameleon-7B and Qwen2-VL-7B-Instruct, representing VLMs of two architectures. Extensive experiments conducted on M^{3}CoT, ScienceQA, and LLaVA-W, under both zero- and few-shot scenarios, have proven that ICoT achieves substantial performance (up to 14%) compared with the existing multimodal CoT methods. Additionally, in the analysis section, we conduct a preliminary exploration of implementing ICoT by copying the KV cache of optimal visual tokens and explain the inner trade-off between efficiency and performance in this approach.

Although ICoT has proven its effectiveness in this paper, we consider ICoT still has significant potential for further improvement. The patch selection in ADS requires storing attention scores, which brings additional memory overhead. Moreover, the fixed number of selected patches in the ADS design is sub-optimal, resulting in unexpected outputs for VLMs. To address these, we intend to incorporate established techniques from segmentation or grounding methods to create a more robust implementation of ICoT. In the future, we also plan to evaluate it across additional backbones and benchmarks to better assess its generalization ability.

7. Acknowledgement

I would like to express my sincere gratitude to all the authors and reviewers for their valuable contributions to this research. This work was supported by the National Natural Science Foundation of China (NSFC 62106165) and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions, China.

References

[1] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 3

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 3

[3] BIG bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. 1

[4] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

[5] Claus Bundesen. A theory of visual attention. *Psychological review*, 97(4):523, 1990. 2

[6] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*, 2024. 1, 3, 5

[7] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024. 2

[8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023. 1

[9] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 1

[10] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *arXiv preprint arXiv:2406.11832*, 2024. 3

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[12] Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*, 2024. 1, 3, 5, 6

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3

[14] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 3

[15] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 3, 5

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[17] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*, 2021. 1

[18] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024. 2, 3, 4

[19] 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. 1, 3, 5, 6

[20] OpenAI. Learning to reason with llms, 2023. 1

[21] Michael I Posner, Steven E Petersen, et al. The attention system of the human brain. *Annual review of neuroscience*, 13(1):25–42, 1990. 2

[22] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 1

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 1

[24] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022. 3, 5

[25] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2, 3, 4, 6

[26] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022. 1

[27] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang,

Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 3

[28] Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19162–19170, 2024. 1, 3

[29] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 3, 6

[30] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 2, 3, 4

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

[32] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 3

[33] Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024. 1, 5, 6

[34] 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. 1, 3, 5, 6

Interleaved-Modal Chain-of-Thought

Supplementary Material

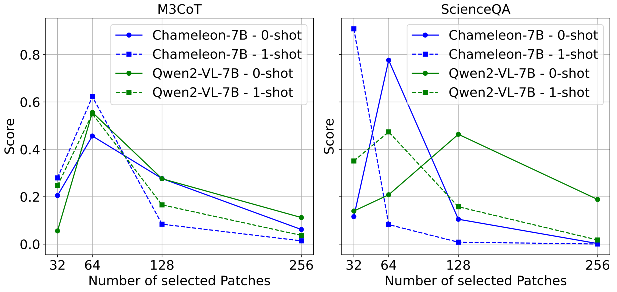


Figure 4. The results of ICoT across validation sets of two datasets on both Chameleon and Qwen2-VL, with the number of selected patches set to 32, 64, 128, and 256. The reported scores are normalized for simplicity.

8. Analysis on the Selected Patches

Intuitively, the performance of ICoT is sensitive to the number of selected patches. If ADS selects a large number of patches every time, the selected patches will be dispersed, resulting in more noise introduced and higher computation costs. In contrast, only a few selected patches perhaps failed to contain enough fine-grained visual information. It is nontrivial to determine the exact number of patches selected by ADS, as fine-grained information in an image is not always the same size. Therefore, in Figure 4, we empirically set the number of patches selected by ADS n to 32, 64, 128, and 256 at a coarse-grained level and illustrate their performance variance across two benchmarks [3]. Observed results indicate that setting n too large or too small is not good for VLMs, and ICoT achieves relatively better performance when n is set to 64.

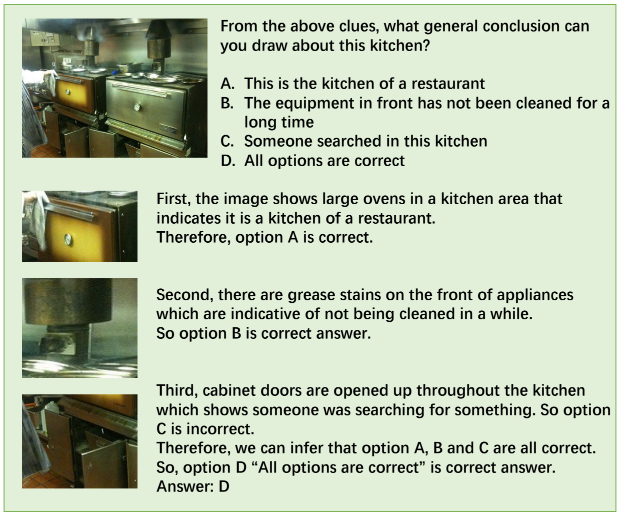
9. Performance on General Benchmark

1-shot	Flickr30k (CIDEr ↑)	OKVQA (VQA-ACC ↑)
Chameleon	22.3	26.2
+ICoT	23.6	28.2

Table 5. Evaluation on general benchmarks

ICoT is a plug-and-play prompting method designed for complex multimodal reasoning, while the performance of ICoT on tasks requiring weak reasoning ability is still unknown. To explore whether ICoT causes degradation, we evaluate ICoT on captioning and VQA in Tab.5. Results indicate advantages of ICoT.

[3] LLaVA-W only contains a test set.



From the above clues, what general conclusion can you draw about this kitchen?

A. This is the kitchen of a restaurant
B. The equipment in front has not been cleaned for a long time
C. Someone searched in this kitchen
D. All options are correct

First, the image shows large ovens in a kitchen area that indicates it is a kitchen of a restaurant.
Therefore, option A is correct.

Second, there are grease stains on the front of appliances which are indicative of not being cleaned in a while.
So option B is correct answer.

Third, cabinet doors are opened up throughout the kitchen which shows someone was searching for something. So option C is incorrect.
Therefore, we can infer that option A, B and C are all correct.
So, option D “All options are correct” is correct answer.
Answer: D

Figure 5. The case of demonstration with Fine-grained Visual Information (FVI), which is used in 1-shot ICoT.

10. Detail Declaration

In Fig. 5, we provide a case to illustrate the FVI in 1-shot ICoT. In Algorithm 1, the stopping criteria is maximum generation length or generating the special token of “end of sequence”.