Pattern Recognition 180 (2026) 114570

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Boosting Multimodal Chain of Thought Reasoning by Selective Mixture of Experts☆

Qilei Li [a], Shitong Sun [b], Da Li [c,b,*], Timothy Hospedales [d], Shaogang Gong [b]

[a] Laboratory for Artificial Intelligence and New Forms of Education, Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan, 430079, China
[b] Queen Mary University of London, London, E1 4NS, United Kingdom
[c] University of York, York, YO10 5DD, United Kingdom
[d] University of Edinburgh, Edinburgh, EH8 9YL, United Kingdom

ARTICLE INFO

Keywords:
Multimodal reasoning
Chain of Thought
Mixture of Experts

ABSTRACT

Extending Language Models (LMs) with a Chain of Thought (CoT) for multimodal reasoning, such as visual question answering, is emerging. Existing methods tried to integrate and interact visual embeddings with text tokens plainly, such as by a simple cross-attention between raw visual and text embedding. We argue that this design limits the LM's ability in comprehending visual input and extracting key information from vari-field input. Consequently, this can cause humiliation in generated rationales. Alternatively, we propose a learnable SElective Mixture of Experts (SEME) module, which plays as a mixture of (domain) experts that (i) *interpret* the visual embedding to LM from *multiple* perspectives to bridge the modality gap from various viewpoints. (ii) *summarize* the *most* relevant knowledge to form a more discriminative embedding, and facilitate the language model (LM) in understanding the visual content and deducing better intermediate rationales and thus more accurate answers. SEME is plug-and-play and readily improves the existing multimodal CoT reasoning models, such as MMCoT and DDCoT. Experimental results on ScienceQA and AOKVQA benchmarks demonstrate the competitive performance of SEME, with just a small LM with 276M parameters, compared with the zero-shot multimodal LLMs, such as GPT4 with more than 175B parameters.

1. Introduction

Large language models (LLMs) [1] have shown excellent reasoning ability in natural language processing. However, limited by their high computational cost and inference inefficiency, recent research has also drawn attention to learning compact LM models (with less than 1B parameters) for specific tasks by instruction tuning. These models also demonstrated high reasoning ability for NLP tasks, especially when elicited with optimized prompting techniques, such as Chain of Thought (CoT) [2,3], which mimics the multi-step reasoning process of human cognition. Despite the outstanding performance in NLP, LMs are limited by their uni-modal reasoning abilities, which fall short for complex tasks involving queries from multiple modalities, such as vision and language, the most common use case in the real world. Thus, recent efforts have increasingly aimed to extend LMs' CoT reasoning to multimodal tasks, *e.g.,* knowledge-based visual science question answering [4,5].

Numerous approaches have been proposed to this end [6–14]. These multimodal CoT based models typically adopt a dual-stage pipeline consisting of a rationale generation followed by an answer inference. Representatively, MM-CoT [6] pioneered the CoT process into multimodal reasoning. They reveal that purely only textual input (*e.g.,* image caption + question) lead LM to produce hallucinated rationales for multimodal reasoning [15]. They thus integrate the visual features with questions as the fused input fed into the LM to generate sensible rationales, which are then used in addition to input images and questions for the final answer inference. DD-CoT [7] further proposed to break the vanilla input questions into multiple sub-questions by another LLM, *e.g.,* ChatGPT3.5, where each sub-question has distinct duties in reasoning the rationale. KAM-CoT [8] exploited a pretrained graph encoder to build the prior knowledge graph to facilitate the LM's comprehension and reasoning of complex tasks. T-SciQ [9] employed ChatGPT3.5 as a teacher model to generate rationales and then use them as the supervisor to facilitate the training of rationale generation network. Recent work such as Corvid [11], retrieval-augmented CoT [12], MedCoT [13], and routing-based experts [14] has further advanced multimodal reasoning from complementary perspectives. With

☆ This article is part of a Special issue entitled: 'PR_Heterogeneous Data Fusion' published in Pattern Recognition.

* Corresponding author.
E-mail address: dali.academic@gmail.com (D. Li).

https://doi.org/10.1016/j.patcog.2026.114570
Received 6 March 2026; Received in revised form 11 July 2026; Accepted 27 July 2026
Available online 6 August 2026

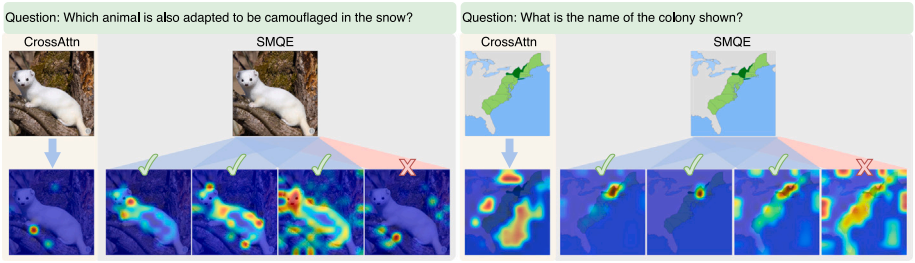


Fig. 1. The visualization of attention maps between image patches and text tokens from different methods. CrossAttn denotes the MM-CoT [6] baseline with plain cross-attention, while SEME denotes the model integrated with the proposed module. Warmer regions indicate higher attention responses. Compared with CrossAttn, SEME focuses more on question-relevant visual regions and provides more interpretable multimodal reasoning.

the guidance of such extra priors, these methods exhibited improved performance in inferencing a more logical rationale compared with the vanilla LM, therefore achieved improved reasoning.

However, existing multimodal CoT methods typically use a single cross-attention operation to fuse raw visual and textual embeddings. Although this operation provides text-conditioned weighting, it does not explicitly route different image–question pairs to specialized knowledge. The comparison in Fig. 1 illustrates this limitation: MM-CoT distributes attention beyond the regions needed for the current question. This observation motivates us to formulate multimodal fusion as a question-conditioned expert-selection problem.

To resolve this issue, inspired by the recent advancement of the Mixture of Experts (MoEs) in LLM, in this work, we propose a SElective Mixture of Experts (SEME) module to improve the multimodal knowledge understanding and association. SEME plays as a lightweight mixture of (domain) experts fulfilling two goals: (i) *interpret* the visual embedding to LM from *multiple* perspectives to bridge the modality gap from various viewpoints. (ii) *summarize* the most relevant multimodality representations to create a more discriminative multimodal embedding, which can facilitate LM in deducing better intermediate rationales and thus more accurate answers.

Generally, SEME aims to *selectively* interpret the visual input from diverse perspectives and associate the most relevant aspects with language embeddings. Specifically, SEME consists of a Gating Transformer (GT) and a Mixture of Query Experts (MQE) layer. The GT layer jointly considers the given image and question tokens and output the corresponding scores for all candidate query experts. Then, experts with Top-K scores will be activated to individually deliver expertise into the visual embedding and fuse with text tokens, so to enhance the intelligibility of the fused multimodal embedding, and makes it more interpretable by the language model (LM). The multimodal embedding will be decoded into a rationale during the first stage, which is then concatenated with the image and question tokens to be processed by SEME again. In the second stage, the LM will infer the final answer to the question based on the input image and the generated rationales. Intuitively, the experts know where to look in an image for certain types of questions to discover the answers. This is important for visual QA, which requires an understanding of the different types of questions and where to find the right clues in the input image for answering. Empirically, we found these learnable query experts are also connected with certain categories of questions, showing the understanding of different knowledge when fused in the LM inference. These connections enable the LM to generate the proper rationales and subsequently accurate answers.

Our contributions are summarized as follows:

(i) We formulate visual–text fusion for multimodal CoT as question-conditioned expert selection and use attention visualizations to identify the need for specialized visual processing.

(ii) We propose a SElective Mixture of Experts (SEME) module with conditional gating and query experts. It adaptively selects question-relevant experts and enhances visual representations from multiple knowledge perspectives before cross-modal fusion.

(iii) We design an expert-enhanced fusion strategy that integrates selected visual expertise with text tokens, producing more discriminative multimodal representations for grounded rationale generation and reliable answer prediction.

(iv) We integrate SEME into representative multimodal CoT frameworks, including MM-CoT and DD-CoT, and demonstrate consistent improvements on ScienceQA and A-OKVQA with a compact language model.

2. Related work

2.1. Multimodal CoT reasoning for LLM

Chain of thought (CoT) [3] refers to an optimized prompt engineering that simulates a human's cognitive thinking process. Recently, CoT [2,16] has significantly improved an LLM's reasoning by prompting them to think step by step. Followups tried to optimize reasoning by breaking down questions into several sub-questions [17]. Due to its effectiveness, CoT has also been extended to multimodal reasoning. UnifedQA [18] first proposed CoT for multimodal reasoning and released a benchmark for evaluation on visual science QA. Different from UnifedQA that generates jointly rationales and answers *at once*, MM-CoT [6] then introduced a two-stage framework *progressively* conducting rationale generation and answer inference. This framework inspired the follow-ups like DD-CoT [7], KAM-CoT [8], and T-SciQ [9], which improved rationales in this two-stage pipeline by leveraging other large-scale pre-trained models, such as ChatGPT3.5. More recently, Corvid [11], retrieval-augmented CoT [12], MedCoT [13], and routing-based experts [14] have advanced multimodal reasoning from complementary perspectives. While these multimodal CoT methods can enhance reasoning performance, their multimodal knowledge aggregation strategies remained plain, such as a basic cross-attention between raw visual and text features. These methods fail to extract crucial knowledge from varied-filed questions, thus limiting effective reasoning. In this work, we propose a selective mixture of query experts module to explore specialized knowledge from the multimodal inputs of different topics, thereby learning a more discriminative multimodal representation for answering a given question.

Overall, existing multimodal CoT methods show the value of intermediate rationales for visual-language reasoning. However, many of them still rely on direct fusion between raw visual and textual embeddings, which may introduce redundant visual cues and provide insufficient grounding for rationale generation. This limitation motivates us to design a selective expert-based module that adaptively extracts question-relevant visual knowledge before rationale and answer generation.

2.2. Large vision-language models

Meanwhile, significant effort has recently been devoted to developing large vision-language foundation models (LVLMs), which connect

visual encoders with large language models for image-text understanding, visual question answering, image captioning, and multimodal dialogue. Representative works include BLIP-2 [19], which introduces a Q-Former to bridge frozen image encoders and frozen LLMs. Besides, InstructBLIP [20], which further improves multimodal instruction-following ability through instruction tuning. And LLaVA [20], which constructs visual instruction-following data to enhance general multimodal dialogue and reasoning. Although these LVLMs have shown strong multimodal understanding ability, they usually rely on large-scale vision-language pretraining, massive instruction data, and billions of parameters, which differs from the compact LM-based multimodal CoT setting studied in this work. Moreover, most LVLMs mainly focus on general vision-language alignment, while question-specific visual evidence selection for rationale generation is less explicitly explored. In contrast, SEME is designed as a lightweight expert module for existing multimodal CoT frameworks. It adaptively selects question-relevant visual expertise and produces more discriminative multimodal representations for rationale generation and answer prediction. The Q-Former in BLIP-2 is related to SEME because both use learnable queries as a bridge between visual and textual information. However, the Q-Former mainly aims to align image embeddings with LLMs through large-scale pretraining, whereas SEME uses conditional gating to activate a sparse subset of query experts for each input, which supports sample-adaptive knowledge selection for multimodal CoT reasoning.

2.3. Mixture of Experts (MoEs)

The concept of MoEs [21] was originally proposed to fuse a few separate models specializing to different tasks to result in a model with enhanced capacity. In MoE, multiple models, known as "experts", are trained to specialize in different aspects of the data. Sparse Mixture of Experts (Sparse MoE) is an extension where only a subset of the experts is activated for any given input. This sparsity significantly reduces the computational burden, which enables Sparse MoE to scale efficiently for large models and datasets. By activating a subset of experts, sparse MoE achieves high performance while optimizing resource usage. Related work includes hierarchical experts [13] and dynamic expert routing [14] in multimodal reasoning. In SEME, we design query experts to model diverse knowledge fields, and a gating transformer to *selectively* activate specific experts, thereby enhancing the understanding for input across various domains.

3. Methodology

3.1. Preliminary: Multimodal CoT Reasoning

Multimodal CoT Reasoning simulates the human's cognitive process by making models to reason step by step. The input is typically a multimodal pair $\boldsymbol{X} = \{\boldsymbol{x}_{\text{text}}, \boldsymbol{x}_{\text{img}}\}$, where $\boldsymbol{x}_{\text{text}}$ is a text description composed of a question, choices, and optional context, and $\boldsymbol{x}_{\text{img}}$ is an image input to provide visual information. Instead of directly predicting the answer based on the input, multimodal CoT approaches [6–8] generally follow a two-stage pipeline, in which there is firstly a rationale reasoning model $\mathbf{F}_{\text{rationale}}(\cdot)$ applied as:

$$\boldsymbol{r} = \mathbf{F}_{\text{rationale}}(\boldsymbol{x}_{\text{text}}, \boldsymbol{x}_{\text{img}}), \tag{1}$$

where $\boldsymbol{r}$ is the reasoned rationales from an LM. It is then combined with the original input to form a new text input as $\boldsymbol{x}'_{\text{text}} = [\boldsymbol{x}_{\text{text}}; \boldsymbol{r}]$. Subsequently, an answer inference model $\mathbf{F}_{\text{answer}}(\cdot)$ applied to make the predicted answer $\boldsymbol{a}$ based on the new input as

$$\boldsymbol{a} = \mathbf{F}_{\text{answer}}(\boldsymbol{x}'_{\text{text}}, \boldsymbol{x}_{\text{img}}). \tag{2}$$

3.2. Mixture of query experts learning

To ensure a logical rationale produced by the LM while reducing hallucinations, we design a mixture of query experts (MQEs) for knowledge specialization and integration. The MQEs are designed in a straightforward manner: Starting from the non-local interaction ability of cross-attention, we initialize a set of queries, each optimized iteratively by associating and archiving specific knowledge with input training samples. Once trained, these experts can adaptively assess the relevance of novel visual–text samples, and then selectively activate the most experienced query experts to contribute prior knowledge as guidance. This is conceptually essential to implicitly build a graph of knowledge that facilitate the LM in conducting reasoning inference.

Specifically, we randomly initialize a set of learnable queries $\boldsymbol{E} = \{\mathbf{e}_i\}_{i=1}^{N}$, each $\mathbf{e}_i \in \mathbb{R}^{d_s}$ termed as an expert to be optimized. These query experts will capture intrinsic properties of the input. Given an visual embedding $\mathbf{v} \in \mathbb{R}^{d_v}$, it will be associated with each expert over T steps to achieve interactive optimization. For notation simplicity, we ignore the expert index i in the following formulations. Specifically, In tth step, the previous query expert $\mathbf{e}_{(t-1)}$ and the input $\mathbf{v}$ are transformed into a shared space followed by a dot-product attention. This aims to interact with certain parts of the input features with experts to update $\mathbf{v}_t$, which is formulated as:

$$\boldsymbol{M} = \frac{1}{\sqrt{d_s}} \mathbf{F}_K(\mathbf{v}) \cdot \mathbf{F}_Q(\mathbf{e}_{t-1}) \in \mathbb{R}^{d_v \times d_s}, \tag{3a}$$

$$\text{attn}_{i,j} = \frac{\exp(M_{i,j})}{\sum_l \exp(M_{i,l})}, \quad \boldsymbol{W}_{i,j} = \frac{\text{attn}_{i,j}}{\sum_{l=1}^{N} \text{attn}_{l,j}}, \tag{3b}$$

$$\mathbf{v}_t = \boldsymbol{W} \cdot \mathbf{F}_V(\mathbf{v}) \in \mathbb{R}^{d_v}, \tag{3c}$$

$$\mathbf{e}_t = \mathbf{e}_{t-1} + \text{MLP}\left(\text{LayerNorm}\left(\text{GRU}(\mathbf{e}_{t-1}, \mathbf{v}_t)\right)\right) \in \mathbb{R}^{d_s}. \tag{3d}$$

where $\mathbf{F}_Q(\cdot)$, $\mathbf{F}_K(\cdot)$, and $\mathbf{F}_V(\cdot)$ stands for linear projectors shared by all the experts to reduce the number of parameters to be learned. $\text{GRU}(\cdot)$ is a Gated Recurrent Unit [22], and $\text{MLP}(\cdot)$ stands for a Multi-Layer Perceptron. After iteratively repeating this process over T times, the optimized visual representation $\mathbf{v} \in \mathbb{R}^{d_v}$ has been enhanced the expertise from $\mathbf{e}$, which, in return, will update the expertise of $\mathbf{e}$ during model optimization.

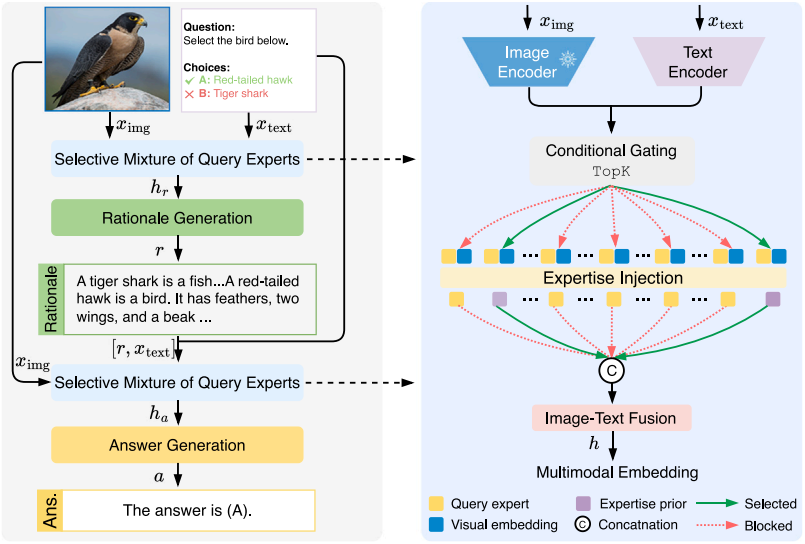
3.3. SElective Mixture of Experts (SEME)

Further to MQEs, we propose the SEME module to achieve sparsely *selective* knowledge interaction and aggregation. The architecture is modular as illustrated in Fig. 2. SEME aims to capture multi-faceted essential details from both modalities to facilitate the language model (LM) in deducing better intermediate rationales and thus more accurate answers. SEME module is designed for plug-and-play integration and can be trained end to end. Therefore, it can be seamlessly incorporated into existing multimodal CoT frameworks.

3.3.1. Conditional gating

Interacting visual representation with each expert can assess the relevance based on the intrinsic properties to each expert. However, *not all pieces of information are beneficial for answer inference*. Introducing an excessive amount of redundant or irrelevant information can distort the visual embedding representation and result in model collapse. To address this issue, it is essential to highlight only the key critical aspects of knowledge with criteria that are determined based on the consensus of the multimodal inputs. Therefore, we propose to selectively activate only dependent experts with a gating network $\mathbf{F}_{\text{gating}}(\cdot)$, which functions as an indexer. The activation of this gating network relies on the interaction between the vision and language modalities.

To this end, we derive a gating network to model the long-range dependency between text token $\boldsymbol{t} \in \mathbb{R}^{d_t}$ and visual embedding $\boldsymbol{v} \in \mathbb{R}^{d_v}$. Specifically, to this end, the hidden dimension of the image embedding is aligned towards the text embedding and then concatenated along the



(a) SMQE for Multimodal CoT reasoning (b) Selective Mixture of Query Experts (SMQE)

Fig. 2. Illustration of the integrating proposed *SElective Mixture of Experts* (SEME) module to a multimodal CoT model. It is a two-stage CoT model benefits from the discriminative representation produced by the SEME module. which consists of a gating transformer and a mixture of expert layers. The gating transformer adaptively selects the experts with the highest relevance (Top-K). The selected experts contribute expertise to fuse with text tokens, thereby creating a better representation.

channel dimension. We use a multi-head attention layer for correlation mining with a learnable selection token [tok] prepend into the joint-modality representation. Meanwhile, a 1D standard learnable positional embedding is adopted to retain positional information for both patch-wise visual representation and token-wise text representation. The output $\boldsymbol{c}$ is sent to the scoring head to determine the activation scores corresponding to each expert. The function of the gating network $\mathbf{F}_{\text{gating}}(\boldsymbol{v}, \boldsymbol{t})$ is composed of the following processes:

$$\boldsymbol{v}' = \text{Projection}_{\text{vt}}(\boldsymbol{v}, d_t) \in \mathbb{R}^{d_t}, \tag{4a}$$

$$\boldsymbol{c} = \text{Correlation}\left(\texttt{tok}, \text{Concat}(\boldsymbol{v}', \boldsymbol{t})\right) \in \mathbb{R}^{d_t}, \tag{4b}$$

$$\boldsymbol{w} = \text{Scoring}(\boldsymbol{c}) \in \mathbb{R}^{N}. \tag{4c}$$

where $\text{Projection}_{\text{vt}}(\cdot)$ represents a linear layer that maps the dimension of $\boldsymbol{v}$ to match $\boldsymbol{t}$, and $\text{Correlation}(\cdot)$ is a multi-head attention layer applied to the inputs, for which the selection token [tok] is projected as the query, and $\text{Concat}(\boldsymbol{v}', \boldsymbol{t})$ serves as the key and value. The scoring head $\text{Scoring}(\cdot)$ is an MLP that produces N-way logits corresponding to each query expert. Upon acquiring the scores $\boldsymbol{w} \in \mathbb{R}^N$ representing per-sample relevance to the N experts, a subsequent routing process is applied based on Top-K scoring as:

$$\mathcal{I}_K = \text{TopK}(\boldsymbol{w}, K), \qquad \boldsymbol{V}_e = \text{MQE}\left(\boldsymbol{v}, \{\boldsymbol{e}_i\}_{i \in \mathcal{I}_K}, T\right) \in \mathbb{R}^{K \times d_v}, \tag{5}$$

where $\mathcal{I}_K$ is the index set of the K largest entries in the gating-score vector $\boldsymbol{w}$, and $\boldsymbol{V}_e$ collects the corresponding K expert-enhanced visual representations. $\text{MQE}(\cdot)$ denotes the iterative expert operation defined in Eq. (3); T is its number of refinement steps; and $K \ll N$ controls the sparsity of expert activation.

It should be noted that the Top-K routing strategy is sample-adaptive. For each input sample, only the experts with the highest gating scores are activated and used to construct the expert-enhanced visual representations ($\boldsymbol{V}_e$). These selected experts participate in the forward computation and are updated through the downstream rationale generation or answer inference loss. The unselected experts are bypassed for this specific sample and do not receive gradients from it. However, this does not mean that they are excluded from training. Since the gating scores depend on the input image-text pair, an expert that is not selected for one sample can still be selected by other samples during training. In this way, different experts are optimized on different subsets of multimodal inputs, which encourages sparse and sample-specific expert specialization. Although this sample-adaptive routing alleviates the concern that unselected experts are completely excluded from training, it does not explicitly guarantee balanced expert utilization. Similar to sparse MoE models, expert usage imbalance may still occur when the gating network repeatedly favors a small subset of experts. In this work, we do not introduce an additional load balancing loss, and we leave explicit expert balancing regularization as a future optimization direction.

3.3.2. Image-text cross modalities fusion

After being optimized by the selected experts, these multi-faceted visual representations will merge with the text embedding via a single-head attention layer:

$$\boldsymbol{t}^v = \mathbf{F}_{\text{attn}}(\boldsymbol{t}, \boldsymbol{V}_e) = \sigma(\boldsymbol{q}_t \boldsymbol{k}_v^T) * \boldsymbol{v}_v \in \mathbb{R}^{d_t}, \tag{6}$$

where $\mathbf{F}_{\text{attn}}(\cdot)$ represents dot-attention function, the query $\boldsymbol{q}_t$ is projected from $\boldsymbol{t}$, the key $\boldsymbol{k}_v$ and value $\boldsymbol{v}_v$ are derived from $\boldsymbol{V}_e$. Subsequently, a gated fusion module [23] is applied to combine the vision and language embeddings and yield the multimodal representation $\boldsymbol{h} \in \mathbb{R}^{d_t}$ by:

$$\begin{aligned} \boldsymbol{h} &= \mathbf{F}_{\text{fusion}}(\lambda, \boldsymbol{t}, \boldsymbol{t}^v) \\ &= (1 - \lambda) \cdot \boldsymbol{t} + \lambda \cdot \boldsymbol{t}^v \in \mathbb{R}^{d_t}, \\ \text{s.t.} \quad \lambda &= \text{Sigmoid}(\boldsymbol{\theta}_t \boldsymbol{t} + \boldsymbol{\theta}_v \boldsymbol{t}^v) \in \mathbb{R}^{d_t}, \end{aligned} \tag{7}$$

where $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_v$ are learnable transformation parameters in $\mathbb{R}^{d_t \times d_t}$. Therefore, by sequentially optimize the embeddings $\{\boldsymbol{v}, \boldsymbol{t}\}$ through Eq. (4) to Eq. (7), we can formulate the entire operation of the SEME module as $\boldsymbol{h} = \mathbf{F}_{\text{SEME}}(\boldsymbol{v}, \boldsymbol{t})$.

Different from the fusion strategy in MM-CoT, the proposed cross-modal fusion is not directly performed between raw visual embeddings and text tokens. In MM-CoT, the visual features are fused with the textual embeddings through a plain cross-attention layer, where all visual tokens are treated as a general visual source. In contrast, SEME first uses conditional gating to select question-relevant query experts. The selected experts then enhance the visual representation from different knowledge perspectives. Therefore, the attention operation in Eq. (6) is performed over the expert-enhanced visual representations $\boldsymbol{V}_e$ rather than the original visual embedding v. The subsequent gated fusion in

Eq. (7) further balances the original textual representation and the expert-guided visual representation. This design makes the cross-modal fusion more question-aware and reduces the influence of irrelevant visual cues.

3.4. Boosting multimodal CoT Reasoning by SEME

Leveraging the modular design of SEME, it can be seamlessly integrated into existing multimodal CoT frameworks, enhancing the provision of informative and specialized multimodal representations. Let $\mathbf{F}_{\text{visual}}(\cdot)$ denote the frozen image encoder, and let $\mathbf{F}_{\text{LM-Encoder}}(\cdot)$ and $\mathbf{F}_{\text{LM-Decoder}}(\cdot)$ denote the encoder and decoder of the LM used in either the rationale-generation or answer-inference stage. The visual and textual embeddings are computed as:

$$\boldsymbol{v} = \mathbf{F}_{\text{visual}}(\boldsymbol{x}_{\text{img}}), \qquad \boldsymbol{t} = \mathbf{F}_{\text{LM-Encoder}}(\boldsymbol{x}_{\text{text}}), \tag{8}$$

The SEME module then produces the multimodal representation, which is decoded into the stage prediction:

$$\boldsymbol{h} = \mathbf{F}_{\text{SEME}}(\boldsymbol{v}, \boldsymbol{t}), \qquad \boldsymbol{p} = \mathbf{F}_{\text{LM-Decoder}}(\boldsymbol{h}), \tag{9}$$

Here, $\boldsymbol{p}$ denotes $\boldsymbol{r}$ in the rationale-generation stage and $\boldsymbol{a}$ in the answer-inference stage; the two stages use separately trained parameters, as detailed in Section 3.5. The processing pipeline of SEME-CoT is depicted in Alg. 1.

For clarity, Alg. 1 denotes the computation of either stage by SEME-Stage. The procedure composes the visual encoder, LM encoder, gating network, selected MQEs, cross-modal attention, gated fusion, and LM decoder defined in Eqs. (4)–(9). With the parameters of the rationale-generation model, it instantiates $\mathbf{F}_{\text{rationale}}(\cdot)$ in Eq. (1); with the separately trained answer-inference parameters, it instantiates $\mathbf{F}_{\text{answer}}(\cdot)$ in Eq. (2).

Algorithm 1 SElective Mixture of Experts (SEME)-CoT for Multimodal Reasoning

Require: Visual input $\boldsymbol{x}_{\text{img}}$, text input $\boldsymbol{x}_{\text{text}}$
Ensure: Generated rationale $\boldsymbol{r}$ and answer $\boldsymbol{a}$

$\boldsymbol{r} \leftarrow$ SEME-Stage$(\boldsymbol{x}_{\text{text}}, \boldsymbol{x}_{\text{img}})$ ▷ Eq. (1)
$\boldsymbol{x}'_{\text{text}} \leftarrow [\boldsymbol{x}_{\text{text}}; \boldsymbol{r}]$
$\boldsymbol{a} \leftarrow$ SEME-Stage$(\boldsymbol{x}'_{\text{text}}, \boldsymbol{x}_{\text{img}})$ ▷ Eq. (2)
return $\boldsymbol{r}, \boldsymbol{a}$
procedure SEME-Stage$(\boldsymbol{x}_{\text{text}}, \boldsymbol{x}_{\text{img}})$
  $\boldsymbol{v} \leftarrow \mathbf{F}_{\text{visual}}(\boldsymbol{x}_{\text{img}})$, $\boldsymbol{t} \leftarrow \mathbf{F}_{\text{LM-Encoder}}(\boldsymbol{x}_{\text{text}})$ ▷ Eq. (8)
  $\boldsymbol{w} \leftarrow \mathbf{F}_{\text{gating}}(\boldsymbol{v}, \boldsymbol{t})$ ▷ Eq. (4)
  $\mathcal{I}_K \leftarrow \text{TopK}(\boldsymbol{w}, K)$
  $\boldsymbol{V}_e \leftarrow \text{MQE}(\boldsymbol{v}, \{\boldsymbol{e}_i\}_{i \in \mathcal{I}_K}, T)$ ▷ Eq. (5)
  $\boldsymbol{t}^v \leftarrow \mathbf{F}_{\text{attn}}(\boldsymbol{t}, \boldsymbol{V}_e)$ ▷ Eq. (6)
  $\boldsymbol{h} \leftarrow \mathbf{F}_{\text{fusion}}(\lambda, \boldsymbol{t}, \boldsymbol{t}^v)$ ▷ Eq. (7)
  $\boldsymbol{p} \leftarrow \mathbf{F}_{\text{LM-Decoder}}(\boldsymbol{h})$ ▷ Eq. (9)
  return $\boldsymbol{p}$

3.5. Optimization objective and training strategy

The proposed SEME module is trained within the two-stage multimodal CoT framework. Following the two-stage multimodal CoT setting, the rationale generation model and the answer inference model are trained separately. Both stages are formulated as conditional sequence generation tasks and optimized with the standard token-level negative log-likelihood objective.

In the first stage, the model takes the image and the original textual input as input and generates the rationale sequence. Given the ground-truth rationale sequence $r^* = \{r_1^*, \ldots, r_{L_r}^*\}$, the rationale generation loss is defined as:

$$\mathcal{L}_r = -\sum_{i=1}^{L_r} \log P_{\theta_r}\left(r_i^* \mid r_{<i}^*, x_{\text{text}}, x_{\text{img}}\right), \tag{10}$$

In the second stage, the rationale is appended to the original textual input to construct $x'_{\text{text}} = [x_{\text{text}}, r]$. Given the ground-truth answer sequence $a^* = \{a_1^*, \ldots, a_{L_a}^*\}$, the answer inference loss is defined as:

$$\mathcal{L}_a = -\sum_{j=1}^{L_a} \log P_{\theta_a}\left(a_j^* \mid a_{<j}^*, x'_{\text{text}}, x_{\text{img}}\right), \tag{11}$$

Here, θ_r and θ_a denote the trainable parameters of the rationale generation model and the answer inference model, respectively. During training, the rationale generation model is supervised by the ground-truth rationale, and the answer inference model is supervised by the ground-truth answer. During inference, the rationale generated by the first-stage model is appended to the original textual input and then used by the second-stage model to predict the final answer.

4. Experiment

4.1. Experimental settings

4.1.1. Implementation details

We used a Vision Transformer (ViT) [34] pre-trained on ImageNet as the image encoder, with the image size of 384 × 384, and the patch size is set to 32. The vision encoder was frozen during training. We used Flan T5-base model [35] as the LM for both rationale generation and answer inference stages. The model was trained for 50 epochs with the learning rate set to 5e-5. The batch size was set to 8. The number of query experts was set to 256, the dimension of each query was set to 1024, and the number of selection rate was set to 0.1, indicating a sparse setting. The optimization iterations T in Eq. (5) was set to 9. All experiments were implemented by PyTorch and HuggingFace, trained on a single NVIDIA A100 40G GPU.

4.1.2. Datasets and evaluation protocol

We evaluated the proposed method on the ScienceQA [4] and A-OKVQA [5] datasets. The ScienceQA dataset consists of 21,208 multiple-choice questions with multimodal contexts drawn from the science topics, including natural science, social science, and language science. It covers significant domain diversity, encompassing 26 topics, 127 categories, and 379 skills. The A-OKVQA dataset is a knowledge-based visual question-answering benchmark with 25,000 questions requiring extensive common sense and world knowledge. We evaluated it using a multiple-choice protocol aligned with ScienceQA with the validation set, since the test set is hidden. We used accuracy as the evaluation protocol, specifically by comparing the ground truth choice with the final prediction generated by the answer generation network.

4.1.3. Evaluation indicators

For ScienceQA, we report accuracy on eight commonly used evaluation splits. NAT, SOC, and LAN denote Natural Science, Social Science, and Language Science, respectively. TXT, IMG, and NO denote questions with textual context, visual context, and no additional context, respectively. G1-6 and G7-12 denote elementary-level and secondary-level questions. Avg denotes the average accuracy over the full test set. For A-OKVQA, we follow the multiple-choice evaluation setting and report the overall answer accuracy on the validation set. These indicators allow us to evaluate the model from different aspects, including subject type, context modality, grade level, and overall reasoning performance.

4.2. Comparison with SOTA methods

As shown in Table 1, traditional VQA models, including Patch-TRM [24], VisualBERT [25], and ViLT [26], achieve relatively limited performance on ScienceQA. This indicates that conventional visual-language matching is insufficient for complex science reasoning, where the model needs to understand both visual evidence and textual question semantics. Compared with these models, multimodal CoT-based methods [7–9] achieve much higher accuracy, which demonstrates

Table 1
Performance on the ScienceQA benchmark, which includes eight splits: Natural Science (**NAT**), Social Science (**SOC**), Language Science (**LAN**), Textual Context (**TXT**), Visual Context (**IMG**), Context-Free (**NO**), Elementary Levels 1-6 (**G1-6**), and Secondary Levels 7-12 (**G7-12**). The best results are in **bold**, the improvements are in blue.

Model	Size	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Heuristic Baselines										
Human	–	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
VQA Models										
Patch-TRM [24]	90M	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50	61.42
VisualBERT [25]	111M	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
ViLT [26]	113M	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
LLM/LLVM Zero-shot Inference										
LMAdapter [27]	6B	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
InstructBLIP [28]	11B	–	–	–	–	90.70	–	–	–	–
LLaVA [20]	13B	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
Chameleon [29]	175B	81.62	70.64	84.00	79.77	70.80	86.62	81.86	76.53	79.93
$\text{ChatGPT}_{\text{CoT}}$ [29]	175B	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03	78.31
GPT-4 w/ CoT [29]	>175B	85.48	72.44	90.27	82.65	71.49	92.89	86.66	79.04	83.99
LM Finetuning										
UnifiedQA [4]	223M	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82	74.11
MM-CoT [6]	223M	84.06	92.35	82.18	82.75	82.75	84.74	85.79	84.44	85.31
MM-CoT+SEME **(Ours)**	276M	**87.92**	**93.48**	**83.64**	**87.19**	**86.22**	**86.20**	**87.56**	**88.73**	**87.97**
		+3.86	+1.13	+1.46	+4.44	+3.47	+1.46	+1.77	+4.29	+2.66
Extra Prior Guided										
$\text{T-SciQ}_{\text{Base}}$ [9]	223M	91.52	91.45	**92.45**	91.94	90.33	92.26	92.11	91.10	91.75
KAM-CoT [8]	223M	93.21	92.21	90.64	93.21	93.26	91.50	92.51	92.42	92.48
DD-CoT† [7]	223M	93.56	94.83	86.45	93.50	93.11	88.50	92.58	90.90	91.68
DD-CoT+SEME **(Ours)**	276M	**94.94**	**96.18**	88.00	**94.97**	**94.15**	**89.97**	**93.83**	**92.62**	**93.40**
		+1.38	+1.35	+1.55	+1.47	+1.04	+1.47	+1.25	+1.72	+1.72

Table 2
Performance on A-OKVQA benchmark. The best results are in bold.

Model	Accuracy
BERT [30]	32.93
GPT-3 (Curie) [1]	35.07
ViLBERT [31]	49.10
FLAN-Alpaca [32]	47.86
IPVR (OPT-66B) [33]	48.60
MM-CoT [6]	50.57
MM-CoT + SEME **(Ours)**	**63.76**

the importance of intermediate rationale generation for multimodal reasoning.

For LM fine-tuning methods, SEME brings consistent improvements over the corresponding baselines. MM-CoT+SEME improves the average accuracy from 85.31% to 87.97%, with a 2.66% absolute gain. The improvement is observed across all eight ScienceQA splits, especially on TXT, IMG, and G7-12, where the gains are 4.44%, 3.47%, and 4.29%, respectively. When applied to DD-CoT [7], SEME further improves the average accuracy from 91.68% to 93.40%, showing that the proposed module can also benefit stronger CoT frameworks with extra reasoning priors.

As shown in Table 2, MM-CoT+SEME improves A-OKVQA accuracy from 50.57% to 63.76%, an absolute gain of 13.19 percentage points over MM-CoT. The resulting representation provides a more focused visual anchor for accessing and applying the relevant language knowledge during rationale generation. Because the grounded rationale is subsequently provided to the answer stage, this benefit can propagate to the final prediction. By comparison, many ScienceQA questions provide supporting textual context in addition to answer choices, which makes the performance less dependent on connecting implicit knowledge to selectively grounded visual evidence. This difference provides a mechanism-level explanation for why SEME yields a substantially larger gain on A-OKVQA. The fact that MM-CoT+SEME also outperforms BERT, GPT-3, ViLBERT, FLAN-Alpaca, and IPVR further supports its effectiveness for knowledge-intensive multimodal reasoning.

4.3. Ablation studies

4.3.1. Component analysis

We evaluated the effectiveness of the SEME module through progressively integrating it into the MMCoT baseline model, either during the rationale generation stage alone or incorporating it further into the subsequent answer inference stage. The results in Fig. 3 confirm that SEME module significantly enhances performance when used solely for rationale generation. Moreover, further improvement is observed when employed for answer generation, as evidenced by the progressively enhanced accuracy across all eight evaluation protocols. It is worth to notice that the improvement observed in the answer inference stage is not as significant as in the rationale generation stage. This demonstrates that a logical and accurate rationale is crucial for deducing the correct answer. Nevertheless, this also highlights that the SEME module is beneficial in enhancing the quality of multimodal representations.

4.3.2. Selected experts to questions' categories

The underlying correlation of the selected experts to the nature (category) of the question is a fundamental assumption for the SEME-CoT model. We validated this by predicting the question's category based on the indices of adaptively selected experts. In ScienceQA benchmark, the questions across 127 categories, we thereby constructed a 127-way classification task, and employed a random forest model to learn the pattern of the selected indices. We set the number of experts to 256 to ensure a more complete coverage compared to the number of question categories, and evaluated the accuracy on the unseen validation set. As shown in Fig. 4, increasing the number of experts to a certain extent results in a strong correlation with the question categories. This supports our assumption that the indices correlate well with the categories. Additionally, we found that the most balanced number of experts is around 25, therefore we set the sparsity to 0.1 in our final model.

4.4. Visualization

4.4.1. Activation maps from top experts

To intuitively understand the focuses of the selected experts, we visualized the attention maps in Fig. 5 by overlaid the attention maps

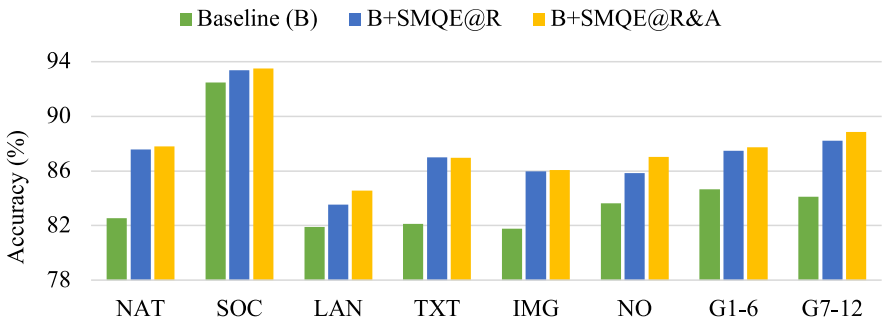


Fig. 3. Analysis of the proposed SEME module. We progressively integrated the SEME module into the baseline and observed consistent enhancements across various protocols. The B denotes the baseline MM-CoT model, B+SMQE@R denotes adding SEME only to the rationale generation stage, and B+SMQE@R&A denotes adding SEME to both the rationale generation and answer inference stages.

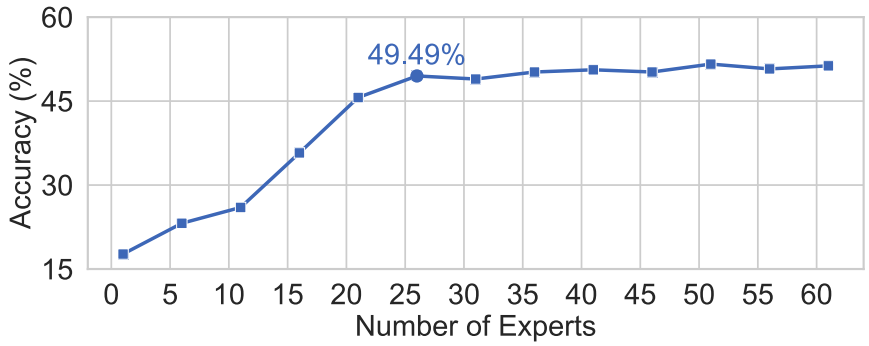


Fig. 4. Impact of a number of experts to question category. It can be observed that an increase in the number of experts to certain extent can achieve higher classification accuracy.

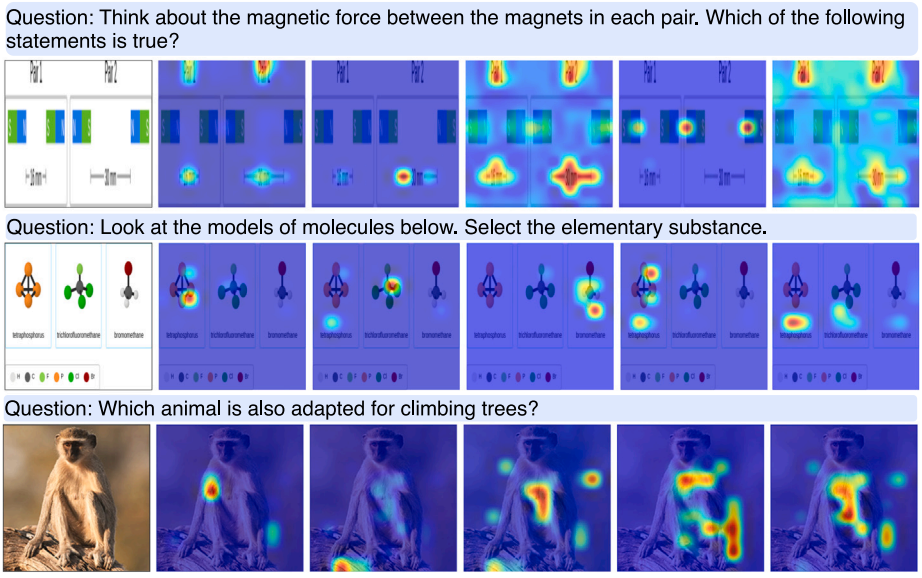


Fig. 5. Attention maps activated by top experts. It reveals the SEME's capacity to focus on distinct aspects by different experts. Warm color indicates high pixel value.

corresponding to Top-K experts over the original images. The visualization indicates that each top expert attends to distinct aspects of the image. For example, with the given image of magnet pairs, we observed that the attention maps mainly focus on instructive regions or the whole objects, especially these corresponding to the correction answer. This pattern is mirrored in the case of the molecules and the monkey queries. Through these examinations, we validated that the proposed SEME module improves the comprehension of the knowledge, and consequently, boost the discrimination of the multimodal representation.

4.4.2. *Comparison of results from different models*

We compared the responses, including the rationale and answer, from different models, including the Flan-T5, MM-CoT, and DD-CoT.

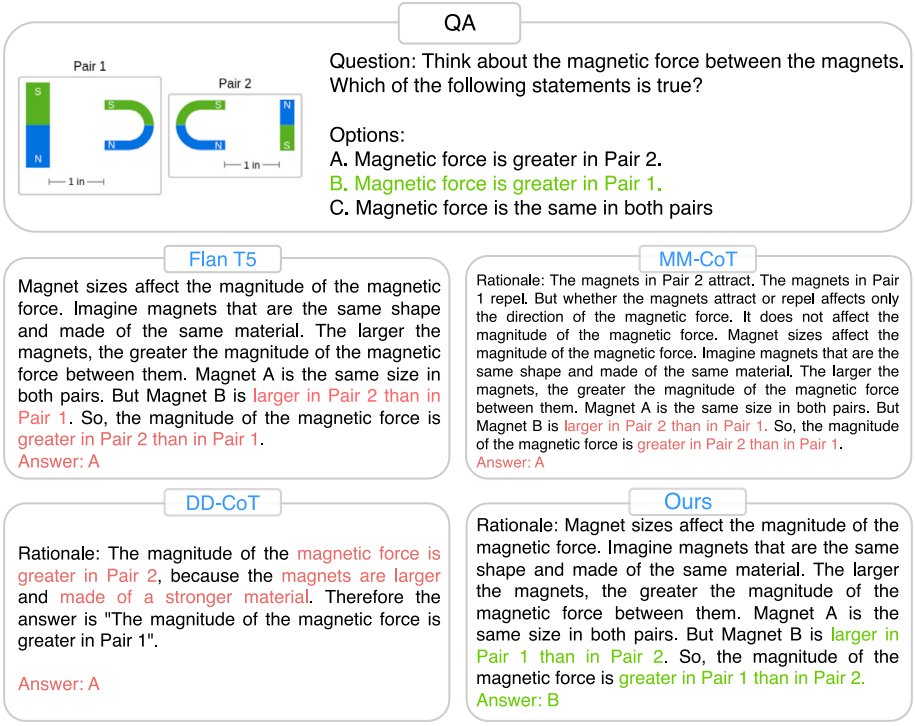


Fig. 6. Illustrative comparisons of Flan T5, MM-CoT, DD-CoT, and proposed SEME models in the context of rationale generation and answer prediction. The correct predictions are highlighted in green, and incorrect predictions are in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Ours model refers to the integration of SEME modules upon the MM-CoT model. One example is shown in Fig. 6. As Flan-T5 is purely an LM, it requires image captions to integrate visual information. Nevertheless, unlike image features, captions lack the ability to convey semantic knowledge of visual content and result in inaccurate rationales. MM-CoT employs a plain cross-attention layer to merge visual information with text embeddings. However, it cannot enrich text tokens with specified knowledge, instead, it degraded to the same performance as Flan T5. DD-CoT generates rationales using ChatGPT3.5 based on the image caption extracted on-the-fly, which suffer the same problem that overlooking semantic visual information. Consequently, it may generate rationales that are correct but irrelevant to the query. In contrast, the proposed SEME-CoT model can generate logical and accurate rationales closely aligned with the query, which is achieved through enhanced multimodal representation and resulted in improved answer accuracy.

5. Conclusion

In this work, we proposed a *SElective Mixture of Experts* (SEME) module, which is designed to bridge visual and language modalities by selectively engaging the most relevant experts for a given question, thereby enhancing the interpretability of a language model (LM) to generate more precise intermediate rationales and, consequently, more accurate answers. It begins with processing image and text embeddings through the gating transformer, which evaluates the correlation and scores potential candidate experts accordingly. The highest-scoring experts, determined by a Top-K selection mechanism, are then processed alongside image embedding through the MoE layer. This integration transforms the query tokens, which are subsequently merged with text tokens within the LM, thereby facilitating the generation of logical rationales. The SEME module is designed as a plug-and-play component, and brings a straightforward enhancement to existing multimodal CoT reasoning frameworks, such as MMCoT and DDCoT. The superior experimental results on ScienceQA and AOKVQA benchmarks demonstrated the competitive performance of SEME with just 276M parameters, compared with the zero-shot multimodal LLMs, such as ChatGPT4 with more than 175B parameters.

Despite these promising results, SEME still has room for further improvement. Since the model follows a two-stage multimodal CoT pipeline, the quality of the generated rationale may affect the final answer prediction. In addition, the current expert selection process is learned implicitly, and its interpretability can be further enhanced. In future work, we will explore more robust rationale generation, more interpretable expert routing, and broader evaluation on complex multimodal reasoning tasks.

CRediT authorship contribution statement

Qilei Li: Formal analysis, Data curation. **Shitong Sun:** Methodology, Investigation. **Da Li:** Supervision. **Timothy Hospedales:** Writing – review & editing, Supervision. **Shaogang Gong:** Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work was in part supported by the China Postdoctoral Science Foundation (Grant No. 2025M781597), the Hubei Provincial Natural Science Foundation of China (Grant No. 2026AFB700), the Fundamental Research Funds for the Central Universities, China (Grant No. XJ2026000901), the Academy of Frontier Interdisciplinary Research at Central China Normal University.

Data availability

Data will be made available on request.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: Proc. of NeurIPS, 2020.

[2] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q.V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, in: Proc. of NeurIPS, 2022.

[3] T. Kojima, S.S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: Proc. of NeurIPS, 2022.

[4] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, A. Kalyan, Learn to explain: Multimodal reasoning via thought chains for science question answering, in: Proc. of NeurIPS, 2022.

[5] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, R. Mottaghi, A-OKVQA: A benchmark for visual question answering using world knowledge, in: Proc. of ECCV, 2022.

[6] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, A. Smola, Multimodal chain-of-thought reasoning in language models, Trans. Mach. Learn. Res. (2024).

[7] G. Zheng, B. Yang, J. Tang, H.-Y. Zhou, S. Yang, DDCOT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models, in: Proc. of NeruIPS, 2024.

[8] D. Mondal, S. Modi, S. Panda, R. Singh, G.S. Rao, KAM-COT: Knowledge augmented multimodal chain-of-thoughts reasoning, in: Proc. of AAAI, 2024.

[9] L. Wang, Y. Hu, J. He, X. Xu, N. Liu, H. Liu, H.T. Shen, T-SciQ: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering, in: Proc. of AAAI, 2024.

[10] L. He, Z. Li, X. Cai, P. Wang, Multi-modal latent space learning for chain-of-thought reasoning in language models, in: Proc. of AAAI, 2024.

[11] J. Jiang, C. Ma, X. Song, H. Zhang, J. Luo, Corvid: Improving multimodal large language models towards chain-of-thought reasoning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 3034–3046.

[12] B. Liu, C. Lyu, Z. Min, Z. Wang, J. Su, L. Wang, Retrieval-augmented multi-modal chain-of-thoughts reasoning for large language models, in: 2025 International Joint Conference on Neural Networks, IJCNN, IEEE, 2025, pp. 1–8.

[13] J. Liu, Y. Wang, J. Du, J.T. Zhou, Z. Liu, Medcot: Medical chain of thought via hierarchical expert, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 17371–17389.

[14] Q. Wu, Z. Ke, Y. Zhou, X. Sun, R. Ji, Routing experts: Learning to route dynamic experts in existing multi-modal large language models, in: The Thirteenth International Conference on Learning Representations, 2025.

[15] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y.J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Comput. Surv. 55 (2023) 1–38.

[16] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, T. Kipf, Object-centric learning with slot attention, in: Proc. of NeurIPS, 2020.

[17] T. Khot, H. Trivedi, M. Finlayson, Y. Fu, K. Richardson, P. Clark, A. Sabharwal, Decomposed prompting: A modular approach for solving complex tasks, in: Proc. of ICLR, 2023.

[18] D. Khashabi, S. Min, T. Khot, A. Sabhwaral, O. Tafjord, P. Clark, H. Hajishirzi, UnifiedQA: Crossing format boundaries with a single QA system, in: Proc. of EMNLP, 2020.

[19] J. Li, D. Li, S. Savarese, S. Hoi, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: Proc. of ICML, 2023.

[20] H. Liu, C. Li, Q. Wu, Y.J. Lee, Visual instruction tuning, in: Proc. of NeurIPS, 2024.

[21] S.E. Yuksel, J.N. Wilson, P.D. Gader, Twenty years of mixture of experts, IEEE Trans. Neural Netw. Learn. Syst. 23 (2012) 1177–1193.

[22] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: Proc. of EMNLP, 2014.

[23] Z. Zhang, K. Chen, R. Wang, M. Utiyama, E. Sumita, Z. Li, H. Zhao, Neural machine translation with universal visual representation, in: Proc. of ICLR, 2019.

[24] P. Lu, L. Qiu, J. Chen, T. Xia, Y. Zhao, W. Zhang, Z. Yu, X. Liang, S.-C. Zhu, IconQA: A new benchmark for abstract diagram understanding and visual language reasoning, in: Proc. of NeurIPS, 2021.

[25] L.H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, VisualBERT: A simple and performant baseline for vision and language, 2019, ArXiv.

[26] W. Kim, B. Son, I. Kim, ViLT: Vision-and-language transformer without convolution or region supervision, in: Proc. of LCML, 2021.

[27] R. Zhang, J. Han, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, P. Gao, Y. Qiao, LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention, in: Proc. of ICLR, 2024.

[28] W. Dai, J. Li, D. Li, A.M.H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. Hoi, InstructBLIP: Towards general-purpose vision-language models with instruction tuning, in: Proc. of NeruIPS, 2023.

[29] P. Lu, B. Peng, H. Cheng, M. Galley, K.-W. Chang, Y.N. Wu, S.-C. Zhu, J. Gao, Chameleon: Plug-and-play compositional reasoning with large language models, in: Proc. of NeurIPS, 2024.

[30] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018.

[31] J. Lu, D. Batra, D. Parikh, S. Lee, ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: Proc. of NeurIPS, 2019.

[32] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T.B. Hashimoto, Alpaca: A strong, replicable instruction-following model, Stanf. Cent. Res. Found. Model. 3 (2023) 7.

[33] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X.V. Lin, et al., OPT: Open pre-trained transformer language models, 2022, 2023, ArXiv.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, 2021.

[35] H.W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, J. Mach. Learn. Res. 25 (2020) 1–53.