Proceedings of the 3rd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2025.25822

C-MMCOT: Multimodal Chain-of-Thought Reasoning Using CLIP Features

Yingche Meng

Independent Researcher, High School Student, Zhengzhou, China
111yoouqian11@gmail.com

Abstract. Chain-of-Thought (CoT) reasoning enhances the performance of large language models (LLMs) on complex tasks such as solving mathematical problems, logical inference, and question answering by guiding models to generate intermediate reasoning steps rather than directly producing final answers. This approach simulates human-like, step-by-step thinking, significantly improving the stability and accuracy of the reasoning process. By moving beyond the “black box" nature of traditional LLM outputs, CoT also lays the foundation for more controllable and multimodal reasoning. However, most existing research has focused on unimodal (text-only) CoT, leaving the multimodal setting underexplored. Multimodal CoT (MMCoT) addresses this gap by separating rationale generation and answer inference through a two-stage architecture that integrates visual and textual inputs. However, due to the limited semantic richness of visual features extracted by the Vision Transformer (ViT), its performance remains suboptimal. In this work, we propose C-MMCoT, a model that leverages CLIP-extracted visual features to generate rationales, thereby enhancing the semantic alignment of visual reasoning. Experiments on the ScienceQA test set demonstrate that C-MMCoT outperforms baseline models. Compared to GPT-4, it achieves higher accuracy on key categories such as SOC, TXT, and IMG, culminating in an overall accuracy that is 0.57 percentage points higher.

Keywords: multimodal reasoning, chain-of-thought, CLIP, ScienceQA, LLMs

1. Introduction

In recent years, large language models (LLMs) [1-4] have demonstrated remarkable capabilities across various natural language processing (NLP) tasks. However, they still face limitations in complex reasoning scenarios [5]. Chain-of-Thought (CoT) [5] reasoning enhances model performance by decomposing a problem into a sequence of intermediate steps, mimicking the human thought process when solving complex tasks. While CoT improves reasoning ability, current research predominantly focuses on unimodal text, which is insufficient for real-world tasks involving intertwined multimodal information [6].

Multimodal reasoning requires integrating different modalities (e.g., images and text) and presents greater challenges than unimodal settings in terms of cross-modal inference, feature extraction, and rationale generation. To address this limitation, the Multimodal Chain-of-Thought (MMCoT) [6] introduces visual information into the CoT framework through a two-stage

DOI: 10.54254/2755-2721/2025.25822

architecture that separates rationale generation from answer prediction. This approach effectively fuses visual and textual features, improving model accuracy and robustness. MMCoT employs Vision Transformer (ViT) [7] for visual encoding. Although ViT excels at extracting discriminative image features, it may lack semantic richness and alignment with language due to limitations in training objectives.

Contrastive Language–Image Pre-training (CLIP) [8], trained on large-scale image–text pairs via contrastive learning, produces semantically rich visual representations that are well aligned with natural language. CLIP features are better at capturing the overall concept and core semantics of an image and exhibit stronger generalisation. To address the semantic shortcomings of ViT-based visual features, we propose C-MMCoT, which utilises CLIP to extract visual features for rationale generation, thereby improving reasoning quality and overall model performance.

2. Related work

2.1. Development of Chain-of-Thought reasoning

With the rise of large language models, Chain-of-Thought (CoT) has become a common strategy for enhancing complex problem-solving. Early work like Auto-CoT [9] introduced the idea of guiding models via reasoning chains. Later approaches, such as SC-CoT [10] used majority voting over multiple chains, while Tree-of-Thoughts (ToT) [11] explored path-searching to improve reasoning accuracy. However, CoT quality depends heavily on model scale. In smaller models, reasoning chains are often redundant or incoherent [12]. As task diversity grows, relying solely on textual input is no longer sufficient for stable and generalizable reasoning.

2.2. Evolution of multimodal reasoning frameworks

To enhance model comprehension, recent research has explored integrating visual information into the reasoning process via multimodal CoT [6,13]. A representative work, MMCoT [6], introduced a two-stage architecture that first generates a rationale and then an answer, enabling visual input to participate in reasoning and improving both interpretability and accuracy. Zheng et al. proposed DDCoT [4], which separates visual recognition into an expert module via role-specific prompting and uses negative-space prompting to let the language model identify uncertain regions that trigger external visual tools. ICoT [14] attempts to jointly generate answers and reasoning in a single step using interleaved image-text chains but lacks explicit control over reasoning quality. In contrast, agent-based systems such as MM-REACT [15] focus more on interaction and tool usage rather than static question answering. This work builds upon the open and representative MMCoT framework, emphasising the relationship between rationale quality and visual features.

3. Methodology

3.1. Overall framework

Our work builds on the Multimodal Chain-of-Thought (MMCoT) framework [6], which adopts a two-stage approach: Stage One generates a textual rationale from the question, options, context, and visual input; Stage Two infers the final answer using the rationale and original inputs.

DOI: 10.54254/2755-2721/2025.25822

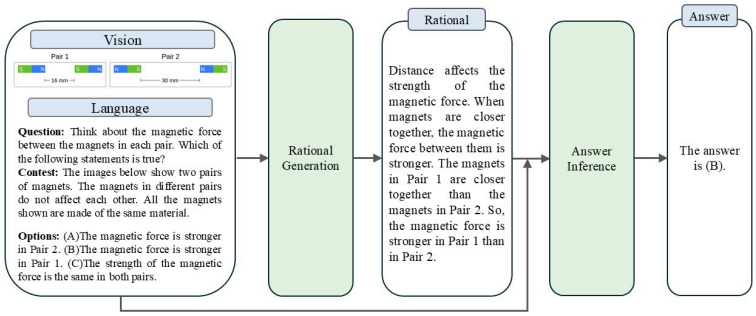

Figure 1. Overall framework of C-MMCoT

Notably, we employ CLIP [8] to generate rationales from image features. Since CLIP embeddings are semantically aligned with language, the model can more effectively associate visual content with contextual and reasoning information. As a result, the generated rationales are more likely to capture key visual cues relevant to the question, rather than merely describing surface-level objects. Moreover, CLIP demonstrates stronger performance in vision-language alignment and abstract relational reasoning. Its representations encode high-level semantic concepts, enabling rationale generation to focus more on abstract reasoning and logical connections beyond low-level pattern recognition.

3.2. Backbone language model

In both stages, we adopt a model based on the T5 encoder-decoder architecture [16] as the backbone. Specifically, we initialise weights from the publicly available declare-lab/flan-alpaca-base model on Hugging Face Hub [17], which is an instruction-tuned variant of FLAN-T5 with 223M parameters. To incorporate visual information, we wrap the base T5 model into a multimodal generation framework (T5ForMultimodalGeneration), centred around a JointEncoder structure designed to fuse visual and textual modalities.

3.3. Visual feature extraction and processing

We adopt CLIP [8] as the visual feature extractor. Following MMCoT [6], we use patch-level image features encoded by the CLIP model with a ResNet-101 (RN101) [18] backbone, resulting in a feature tensor of shape [49, 2048], where each image is represented by 49 patches, each encoded as a 2048-dimensional vector.

In the T5ForMultimodalGeneration model, visual features are first passed through a learnable projection layer to match the embedding space of the language model. The projected visual embeddings are then input into the encoder alongside text embeddings.Within the JointEncoder module, visual features are first projected to the hidden dimension of T5 via image_dense. A multimodal attention mechanism (mha_layer) then uses text hidden states as queries, and the projected visual features as keys and values, to aggregate visual information. Finally, a gating mechanism (gate_dense with sigmoid activation) combines text and visual representations via weighted fusion, constructing the final multimodal representation.

DOI: 10.54254/2755-2721/2025.25822

4. Experiments

4.1. Dataset

We evaluate our model using the ScienceQA benchmark [19], which comprises 21,208 multiple-choice, multimodal questions. The dataset covers diverse domains including natural science, language science, and social science, spanning 26 topics, 127 categories, and 379 distinct skills.ScienceQA is split into 12,726 training samples, 4,241 validation samples, and 4,241 test samples.

4.2. Baseline

We compare C-MMCoT with several representative baselines on the ScienceQA dataset, grouped into three categories:(i) Standard visual question answering (VQA) models, including MCAN [20], BAN [21], DFAF [22], ViLT [23], Patch-TRM [24], and VisualBERT [25];(ii) Language models (LMs), including UnifiedQA [26] (text-to-text), GPT-3.5, ChatGPT, and GPT-4 [27];(iii) Fine-tuned large language models, represented by LLaMA-Adapter [28].

4.3. Implementation details

We follow the MMCoT framework [6] and adopt a T5 encoder-decoder architecture [16], with weights initialised from declare-lab/flan-alpaca-base [17], a public 223M-parameter FLAN-Alpaca model on Hugging Face. Training is conducted on a single NVIDIA RTX 4060 Laptop GPU (8GB) for 15 epochs with a learning rate of 5e-5.

4.4. Results

Table 1 reports the accuracy of final answer generation in Stage Two. Overall, C-MMCoT achieves the best performance across multiple categories. It ranks first on questions involving SOC, TXT, IMG, and G7–12, and second on NAT, NO, and G1–6. The overall accuracy reaches 84.56%, the highest among all models.

Compared to the best-performing VQA model VisualBERT [25], C-MMCoT improves accuracy by 25.44% (NAT), 16.87% (SOC), 21.73% (LAN), 20.53% (TXT), 17.55% (IMG), 27.73% (NO), 21.77% (G1–6), 24.33% (G7–12), and 22.69% on average.Compared to the best LLM, GPT-4 [27], our model underperforms on NAT, LAN, NO, and G1–6, but achieves gains of 13.61% (SOC), 0.59% (TXT), 8.23% (IMG), 5.21% (G7–12), and 0.57% overall.

Against the best fine-tuned LLM, LLaMA-Adapter (T) [28], C-MMCoT shows improvements of 5.77% (NAT), 12.26% (SOC), 2.36% (LAN), 4.91% (TXT), 9.37% (IMG), 3.13% (NO), 4.96% (G1–6), 8.57% (G7–12), and 6.25% on average.Although C-MMCoT has 223M parameters -larger than typical VQA models-it significantly outperforms them. Compared to high-parameter LLMs like GPT-3.5, ChatGPT, and GPT-4, C-MMCoT achieves superior results on most evaluation metrics.

DOI: 10.54254/2755-2721/2025.25822

Table 1. Main results (%) on ScienceQA test set. Classes: NAT =natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. Bolded red indicates the highest accuracy; blue indicates the second-best

Model	Size	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	AVG
MCAN [20]	95M	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72	54.54
BAN [21]	112M	60.88	46.57	66.64	62.61	52.60	65.51	56.83	63.94	59.37
DFAF [22]	74M	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17	60.72
ViLT [23]	113M	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
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
UnifiedQA [26]	223M	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82	74.11
GPT-3.5 [19]	173B	77.71	68.73	80.18	75.12	67.92	81.81	80.58	69.08	76.47
ChatGPT [27]	-	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03	78.31
GPT-4 [27]	-	85.48	72.44	90.27	82.65	71.49	92.89	86.66	79.04	83.99
LLaMA-Adapter (T) [28]	6B	79.00	73.79	80.55	78.30	70.35	83.14	79.77	75.68	78.31
C-MMCoT	223M	84.77	86.05	82.91	83.24	79.72	86.27	84.73	84.25	84.56

5. Conclusion

To address the limited semantic richness of visual features extracted by ViT in MMCoT, we proposed C-MMCoT, a two-stage framework that separates rationale generation from final answer inference. By leveraging CLIP to generate semantically aligned visual features, our model enhances visual reasoning quality. Experiments on the ScienceQA test set show that C-MMCoT outperforms the best VQA model (VisualBERT), the top LLM (GPT-4), and the strongest fine-tuned LLM (LLaMA-Adapter), with average accuracy gains of 22.69%, 0.57%, and 6.25%, respectively.

Notably, our experiments suggest that the ROUGE scores of generated rationales do not always positively correlate with their downstream contribution to answer accuracy. This finding highlights the limitations of text-similarity-based metrics in evaluating the logical soundness of reasoning chains. Future work may explore new evaluation paradigms that directly assess the logical coherence of the reasoning process.

References

[1] Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: Palm 2 technical report. arXiv preprint arXiv: 2305.10403 (2023)

[2] Plaat, A., Wong, A., Verberne, S., Broekens, J., van Stein, N., & Back, T. (2024). Reasoning with large language models: A survey. arXiv preprint arXiv: 2407.11511.

[3] Wang, C., Zhao, J., & Gong, J. (2024). A Survey on Large Language Models from Concept to Implementation. arXiv preprint arXiv: 2403.18969.

[4] Zhou, Z., Ning, X., Hong, K., Fu, T., Xu, J., et al. (2024). A Survey on Efficient Inference for Large Language Models. arXiv preprint arXiv: 2404.14294.

[5] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv: 2201.11903.

[6] Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., & Smola, A. (2023). Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv: 2302.00923.

DOI: 10.54254/2755-2721/2025.25822

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. 9th International Conference on Learning Representations (ICLR 2021), Virtual Event.
[8] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning (ICML 2021), 139, 8748–8763.
[9] Zhang, X., Deng, Y., Jiang, Z., & Rush, A. (2023). Auto-CoT: Automatic chain-of-thought prompting in large language models. International Conference on Learning Representations (ICLR 2023).
[10] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv: 2203.11171.
[11] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems, 36.
[12] Liu, T., Guo, Q., Hu, X., Jiayang, C., Zhang, Y., Qiu, X., & Zhang, Z. (2025). Efficient reasoning with model collaboration. arXiv preprint arXiv: 2504.00424.
[13] Zheng, G., Yang, B., Tang, J., Zhou, H.-Y., & Yang, S. (2023). DDCoT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. Advances in Neural Information Processing Systems, 36, 5168–5191.
[14] Gao, J., Li, Y., Cao, Z., & Li, W. (2024). Interleaved-modal chain-of-thought. arXiv preprint arXiv: 2411.19488.
[15] Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., & Wang, L. (2023). MM–REACT: Prompting ChatGPT for multimodal reasoning and action. arXiv preprint arXiv: 2303.11381.
[16] Raffel, C., Shazeer, N., Roberts, A., Lee, K., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21, 140: 1–140: 67.
[17] Ye, D., Lin, Z., Han, X., Li, J., & Tan, C. (2023, March 21). Flan-alpaca: Instruction tuning for language models with FLAN and Alpaca. GitHub. https: //github.com/declare-lab/flan-alpaca
[18] Luo, J. (2025, January). OpenAI-CLIP-Feature. GitHub. https: //github.com/jianjieluo/OpenAI-CLIP-Feature
[19] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, Ø., Clark, P., & Kalyan, A. (2022). Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems, 35, 2507–2521.
[20] Yu, Z., Yu, J., Cui, Y., Tao, D., & Tian, Q. (2019). Deep modular co-attention networks for visual question answering. CVPR 2019, 6281–6290.https: //doi.org/10.1109/CVPR.2019.00644
[21] Kim, J.-H., Jun, J., & Zhang, B.-T. (2018). Bilinear attention networks. Advances in Neural Information Processing Systems, 31, 1571–1581.
[22] Gao, P., Jiang, Z., You, H., Lu, P., et al. (2019). Dynamic fusion with intra- and inter-modality attention flow for visual question answering. CVPR 2019, 6639–6648. https: //doi.org/10.1109/CVPR.2019.00680
[23] Kim, W., Son, B., & Kim, I. (2021). ViLT: Vision-and-language transformer without convolution or region supervision. Proceedings of the 38th International Conference on Machine Learning (ICML 2021), 139, 5583–5594.
[24] Lu, P., Qiu, L., Chen, J., Xia, T., Zhao, Y., Zhang, W., Yu, Z., Liang, X., & Zhu, S.-C. (2021). IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. NeurIPS 2021 Datasets and Benchmarks Track.
[25] Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). VisualBERT: A simple and performant baseline for vision and language. arXiv preprint arXiv: 1908.03557.
[26] Khashabi, D., Min, S., Khot, T., Sabharwal, A., Tafjord, Ø., Clark, P., & Hajishirzi, H. (2020). UNIFIEDQA: Crossing format boundaries with a single QA system. Findings of EMNLP 2020, 1896–1907. https: //doi.org/10.18653/v1/2020.findings-emnlp.171
[27] Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.-W., Wu, Y. N., Zhu, S.-C., & Gao, J. (2023). Chameleon: Plug-and-play compositional reasoning with large language models. NeurIPS 2023.
[28] Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv: 2303.16199 (2023).