GoCa: Trustworthy Multi-Modal RAG with Explicit Thinking Distillation for Reliable Decision-Making in Med-LVLMs

Pengyu Dai[1], Yafei Ou[1(✉)], Yuqiao Yang[1], Ze Jin[1], and Kenji Suzuki[1]

Institute of Integrated Research, Institute of Science Tokyo, Yokohama, Japan
ou.y.ac@m.titech.ac.jp

Abstract. Medical Large Vision-Language Models (Med-LVLMs) have shown promise in enhancing medical diagnosis by enabling interactive and knowledge-driven healthcare applications. However, these models often suffer from factual hallucinations which may lead to incorrect diagnoses. Retrieval-augmented generation (RAG) has been proposed to mitigate these issues, yet its effectiveness in multi-modal medical applications is hindered by over-reliance on retrieved data and the opacity of text-based reasoning. To address these challenges, we propose GoCa, a multi-modal RAG system based on chain-of-thought (CoT) distillation and explicit thought optimization, which is designed to enhance both the factuality and explainability of Med-LVLMs. Our GoCa consists of three key components: (1) a self-evolving CoT framework that leverages multi-agent collaboration to refine diagnostic reasoning iteratively and (2) a seamless, preference-guided optimization mechanism that distills high-quality CoT reasoning using preference tuning and (3) an adaptive Monte Carlo-like top-k selection strategy. These innovations ensure that the RAG process remains logically transparent and adaptable, significantly improving consistency when integrating retrieve contexts. Experimental results across multiple datasets on medical visual question answering (Med-VQA) demonstrate that GoCa outperforms several recent state-of-the-art methods, achieving superior factual accuracy and coherence. The code can be found at https://github.com/Da1daidaidai/Goca.

Keywords: Vision-Language Models · Retrieval-augmented generation · Chain-of-thought.

1 Introduction

Large Language Models (LLMs) [1,2], especially Large Vision-Language Models (LVLMs), leveraging their powerful multi-modal representational capabilities and advanced reasoning skills, have made significant contributions to various fields, including medical imaging [18,21,28]. However, despite their potential to address clinical challenges in real-world scenarios [24], *The prevalence of reasoning hallucinations* [32], which result in faulty decision-making and inaccurate diagnoses, poses a major obstacle to the deployment and broad adoption of

Med-LVLMs in clinical practice. Given the high-stakes nature of healthcare, even minor diagnostic errors can have serious repercussions for patient care, further underscoring the challenges of integrating these models into real-world clinical scenarios [34,6].

Recent approaches seek to address this issue by factual content enhancement methods like Retrieval-Augmented Generation (RAG) [14]. While optimizing retriever construction improves factual grounding in medical RAG [26], it does not address how Med-LVLMs internally evaluate and utilize retrieved knowledge during reasoning. Thus, beyond improving retrieval mechanisms, it is crucial for Med-LVLMs to independently assess and determine the reliability of factual information. Xia et al. [34] proposed a more consistency-driven RAG-PT (Preference Tuning) strategy [19], which leverages direct preference optimization (DPO) [23] to mitigate instances where incorrect answers persist despite factual enhancement, ensuring alignment between the retriever and the LVLM. Building on this foundation, [33] further refined the approach by addressing finer-grained challenges, such as ensuring alignment across different modalities within the DPO framework. These RAG-PT alignment methods primarily focus on optimizing final outputs. However, they do not enhance explicit reasoning before generating responses. This *absence of explicit reasoning* is particularly problematic for complex questions that demand logical inference and strategic planning [31], as is often required in medical multimodal decision-making.

Orthogonal to external knowledge retrieval, another line of research enhances model reasoning through internal explicit and slow thinking, commonly known as Chain-of-Thought (CoT) [30]. Some studies [7] have explored enhancing the reasoning capabilities of LVLMs by fine-tuning them on instruction sets that explicitly encode medical facts into CoTs. However, constructing such high-quality annotated instruction sets for diverse medical scenarios is both costly and labor-intensive, limiting its scalability. Moreover, multimodal CoT still struggles to understand the nuanced intents behind medical images and textual inputs, limiting its applicability in real-world clinical settings [35]. Recent work on general tasks has recognized the potential of explicit reasoning in preference tuning [31], leveraging an LLM-as-judge approach to reduce reliance on supervisory signals. While in medical tasks, hallucinations become even more problematic when reasoning fails to incorporate accurate medical facts and clinical details, leading to *unfaithful chain of thought* [17]. These aforementioned limitations underscore the need for a more adaptable framework that integrates the factual grounding of RAG-PT with the explicit reasoning capabilities of CoT, while removing the dependency on human-annotated supervision. This leads to the key question this paper aims to answer:

how can we bridge external factual enhancement with explicit internal reasoning in an autonomous manner?

To this end, our primary contribution is **GoCa**, a multi-modal RAG system that leverages chain-of-thought (CoT) distillation and explicit thought-based preference optimization. Our goal is to seamlessly integrate CoT reasoning into the RAG-PT paradigm, shifting Med-LVLM's fact-retrieval and reasoning pro-

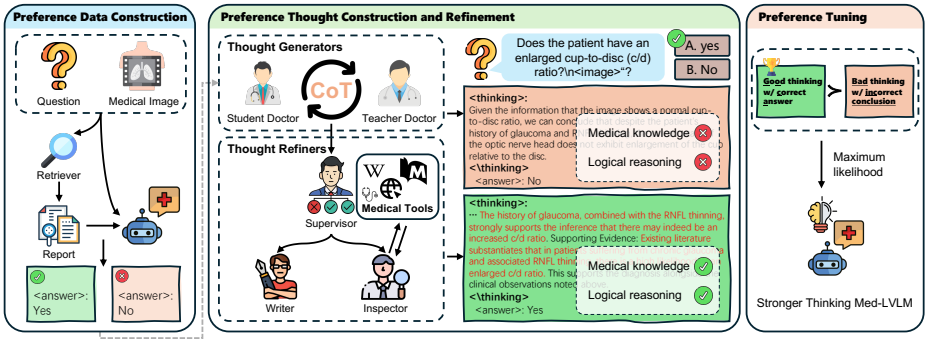

Fig. 1. Overview of the CoCa Workflow: First, the multimodal retriever retrieves reference reports based on the input image. Then, a multi-agent system constructs hierarchical Chains of Thought by leveraging the retrieved reports. Finally, the vision-language model is fine-tuned via preference optimization to fully exploit the CoT data.

cess from a result-oriented approach to a process-driven one. To achieve this, (i) we introduce a **self-evolving multi-agent collaboration framework**, which simulates the formation of diagnostic reasoning in real-world multi-specialist collaborations. This enables the construction of an effective chain of thought that integrates retrieved information with medical knowledge. (ii) This hierarchical collaboration facilitates a **seamless CoT preference optimization strategy**, allowing the LVLM to learn retrieved medical facts from the distilled CoTs and distinguish between *Good reasoning with Correct answers* and *faulty reasoning with incorrect conclusions*. (iii) For retrieval enhancement, we implement a **Monte Carlo-like adaptive Top-k method**, which dynamically explores adaptive top-k selection to refine retrieval effectiveness within the RAG framework. We evaluate GoCa on three medical visual question answering (Med-VQA) datasets. Experimental results demonstrate that GoCa outperforms several recent state-of-the-art approaches, achieving superior factual accuracy and coherence.

2 Methodology

2.1 Context Retrieval for Reference

In this multimodal retrieval phase, GoCa retrieves textual reports most similar to the target image's features, providing image-based medical facts to guide response generation in subsequent phases. Following CLIP [22], the retriever encodes images and reports into embeddings via a vision and text encoder. Specifically, medical images X_{img} are encoded as image representations $V_{img} \in \mathbb{R}^{N \times P}$ using a vision encoder E_{img} (i.e., $V_{img} = E_{img}(X_{img})$), where N denotes the number of medical images and P is the embedding dimension. Likewise, medical reports X_{txt} are encoded into text embeddings $V_{txt} \in \mathbb{R}^{N \times P}$ using a text encoder E_{txt} (i.e., $V_{txt} = E_{txt}(X_{txt})$).

To adapt these encoders to the medical domain, we employ contrastive learning with the loss function shown in Eq. (1), where the similarity matrix $S \in \mathbb{R}^{N\times N}$ is defined as $S = \frac{V_{img}}{|V_{img}|} \cdot \left(\frac{V_{txt}}{|V_{txt}|}\right)^T$. Each element $S_{i,j}$ quantifies the similarity between the image representation of example i and the text representation of example j.

$$L = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{\exp(S_{i,i})}{\sum_{j=1}^{N}\exp(S_{i,j})} + \log\frac{\exp(S_{i,i})}{\sum_{j=1}^{N}\exp(S_{j,i})}\right) \tag{1}$$

2.2 Self-Evolving Medical Explicit Thinking Construction

In this section, we formalize the multi-agent framework for collaborative Chain-of-Thought construction. Given a medical image x_{img}, its corresponding question x_q, answer x_a, and retrieved reports x_{txt} by Section 2.1, our goal is to generate an output CoT $\hat{z}/z$. To achieve this, we decompose the diagnostic process into multiple subtasks, each handled by specialized agents. Specifically, we structure this process into two hierarchical stages.

Thought Generators consist of two specialized agents: a student doctor and a teacher doctor. Given x_q, x_a, and x_{txt}, the objective is to generate a coherent chain-of-thought (CoT) $\hat{z}$ within a maximum of t_1 interaction rounds.

First, the student doctor generates an initial CoT explanation, it can be defined as $z_0 \sim p_\theta(\cdot \mid I^{\text{cot}}_{\text{std}}, x_q, x_a, x_{\text{txt}})$, where p_θ denotes the underlying large language model, and $I^{\text{cot}}_{\text{std}}$ is the prompt guiding the student doctor to derive preliminary reasoning solely from the input.

Next, the teacher doctor evaluates the reasoning and provides corrective feedback, as $d_{t_1} \sim p_\theta(\cdot \mid I^{\text{fb}}_{\text{tch}}, z_0, x_q, x_a, x_{\text{txt}})$. where $I^{\text{fb}}_{\text{tch}}$ instructs the teacher doctor to assess the CoT z_0 and suggest improvements.

In subsequent rounds, the student doctor refines its reasoning by integrating the teacher doctor's feedback along with previous CoT outputs, this process can be defined as $\hat{z}_{t_1} \sim p_\theta(\cdot \mid I^{\text{cot}}_{\text{std}}, x_q, x_a, x_{\text{txt}}, \{z_i, d_i\}_{i=0}^{t_1-1})$. The history $\{z_i, d_i\}_{i=0}^{t_1-1}$ records all prior CoTs and feedback, ensuring each refinement builds upon previous iterations. This process continues until the teacher doctor confirms that the CoT explanation meets diagnostic standards, yielding the final output $\hat{z}$.

Thought Refiners The team comprises three agents: a supervisor, a writer, and an inspector. The supervisor orchestrates the refinement process by dynamically selecting the next agent based on the reasoning state. At each iteration t_2, it determines the action a_{t_2} as $a_{t_2} \sim p_\theta(\cdot \mid I_{\text{sup}}, H_{t_2})$, where I_{sup} is the supervisory prompt, and $H_{t_2} = \{\hat{z}_i, d_i\}_{i=0}^{t_2-1}$ records prior CoT outputs and feedback.

If assigned to the inspector, the agent retrieves relevant biomedical knowledge to verify and enrich the reasoning, as $d^{\text{insp}}_{t_2} \sim p_\theta(\cdot \mid I_{\text{insp}}, \hat{z}_{t_2-1}, x_q, x_a, x_{\text{txt}})$, where I_{insp} guides the retrieval process.

Alternatively, if the writer is activated, it refines the chain-of-thought (CoT) by integrating previous reasoning and inspector feedback, this process can be defined as $z_{t_2} \sim p_\theta(\cdot \mid I_{\text{wrt}}, \hat{z}_{t_2-1}, d_{t_2}^{\text{insp}}, H_{t_2-1})$, with I_{wrt} instructing the writer to generate an improved diagnostic explanation.

This iterative refinement continues until the supervisor issues a `<FINISH>` command, indicating that the CoT explanation meets the required diagnostic standards, at which point the final output z is produced.

2.3 RAG-based Explicit Thinking Distillation with Preference Fine-tuning

Retrieval-augmented Med-LVLM models often generate incorrect answers due to overreliance on retrieved contexts. Previous work [33,34] has addressed this via preference fine-tuning based on correct/incorrect response collection. However, this binary approach lacks interpretability, failing to capture nuanced human judgment, leading to reward hacking issues [25]. Here, we enhance Med-LVLM's reasoning via CoT-based proximal optimization [31], also viewed as reasoning distillation [8] from closed-source LLMs. Specifically, we select samples $\mathcal{D} = \{x_{img}^{(i)}, x_a^{(i)}, x_q^{(i)}\}_{j=1}^N$ from a separate set with samples that are not used to fine-tune the retriever in Section 2.1, where x, x_a, x_q denote input medical image, ground-truth answer, and question, respectively.

We identify responses $a_c = \mathcal{M}(x_{img}, x_q)$ where the model originally answers (i.e., $a_b = x_a$) correctly but gives incorrect answers $a_{ic} = \mathcal{M}(x_{img}, (x_q, x_{txt}))$ after incorporating retrieved contexts as dispreferred responses, as they indicate over-dependence on the retrieval. Conversely, ground-truth answers x_a are considered preferred responses. We define the preference dataset as Eq.(2), where $\mathbf{x}^{(i)} = \left(x_{img}^{(i)}, x_q^{(i)}\right)$ denotes the composite input (medical image and question), $y_p^{(i)} = x_a^{(i)}$ is the preferred (ground-truth) response, and $y_d^{(i)} = a_{ic}^{(i)}$ is the dispreferred response.

$$\mathcal{D}_o = \left\{\mathbf{x}^{(i)}, y_p^{(i)}, y_d^{(i)}\right\}_{i=1}^N \tag{2}$$

Based on this curated preference data, we fine-tune the Med-LVLM via direct preference optimization. Following DPO [23], the loss is computed as

$$\mathcal{L}_{pt_1} = -\mathbb{E}_{(\mathbf{x}, y_p, y_d)\sim\mathcal{D}_o}\left[\log\sigma\Big(\alpha\log\frac{\pi_\theta(y_p \mid \mathbf{x})}{\pi_\theta(y_d \mid \mathbf{x})} - \alpha\log\frac{\pi_\theta(y_d \mid \mathbf{x})}{\pi_\theta(y_p \mid \mathbf{x})}\Big)\right] \tag{3}$$

We then leverage Section 2.2 to generate chain-of-thought outputs for both preferred and dispreferred responses. In our approach, the generated CoT outputs, denoted as $z^{(i)}$ for the preferred response and $\hat{z}^{(i)}$ for the dispreferred response, are prepended with a `<think>` token and integrated into their respective responses. Consequently, we define the augmented dataset as

$$\mathcal{D}_t = \left\{\mathbf{x}^{(i)},\, y_p^{(i)} \oplus z^{(i)},\, y_d^{(i)} \oplus \hat{z}^{(i)}\right\}_{i=1}^N \tag{4}$$

Subsequently, we perform extra round of explicit chain-of-thought preference tuning. The tuning loss is formulated as Eq. (5), where $\sigma(\cdot)$ is the sigmoid function, π_θ represents the reference policy, which is the LLM fine-tuned through supervised learning.

$$\mathcal{L}_{pt_2} = -\mathbb{E}_{(\mathbf{x},\, \hat{y}_p,\, \hat{y}_d)\sim\mathcal{D}_t}\left[\log\sigma\left(\alpha\log\frac{\pi_\theta\left(y_p \oplus z \mid \mathbf{x}\right)}{\pi_\theta\left(y_d \oplus \hat{z} \mid \mathbf{x}\right)} - \alpha\log\frac{\pi_\theta\left(y_d \oplus \hat{z} \mid \mathbf{x}\right)}{\pi_\theta\left(y_p \oplus z \mid \mathbf{x}\right)}\right)\right] \tag{5}$$

2.4 Adaptive Monte Carlo-like top-k Selection

During inference of the RAG system, given a target medical image x_{img} requiring a reference report, we retrieve the top-K most similar medical reports using the method described in Section 2.1. Following [34], we observe that the optimal choice of K varies across different queries. Rather than using a fixed threshold, we employ a Monte Carlo-like approach to dynamically determine the optimal K for each scenario. For a large candidate set $C_K \subseteq \mathbb{N}$, we compute the factuality risk $F_R(k)$ for each k and select the top-p values with the lowest $F_R(k)$, forming a search space $C'_K \subseteq C_K$:

$$C'_K = \{k \mid k \in C_K, \text{rank}_{F_R}(k) \leq p\} \tag{6}$$

The factuality risk is defined as Eq. (7), where x is the target medical image, q is the query, and T_K denotes the top-K retrieved contexts. The function $\text{ACC}(\cdot)$ measures the accuracy of Med-LVLM M, defined as the ratio of correct answers to the total number of answers. We can then iterate over k within the search space C'_K to identify the optimal one. For inference, the C'_K is derived on validation sets to optimize factuality risk, and this search space is then directly employed for the test sets.

$$F_R(k) = 1 - \text{ACC}(M(x, (q, T_K))) \tag{7}$$

3 Experiments

3.1 Experiment Setup

Dataset We employed three publicly available medical vision-language datasets: MIMIC-CXR [12], IU X-ray [4], and HarvardFairVLMed [20], covering radiology and ophthalmology domains. To ensure consistency with prior work and facilitate fair comparative evaluation, we adopted the same dataset structure and same splitting as proposed in [34]. Specifically, the training set was divided into two parts: one for training the retriever and the other for constructing the preference dataset used in Section 2.3. Constructing validation data involved processing reports from the preference dataset and the original test set using GPT-4 [1] to generate closed-ended visual question answering (VQA) pairs and each question was designed in a yes/no format to ensure concise and fact-based evaluation.

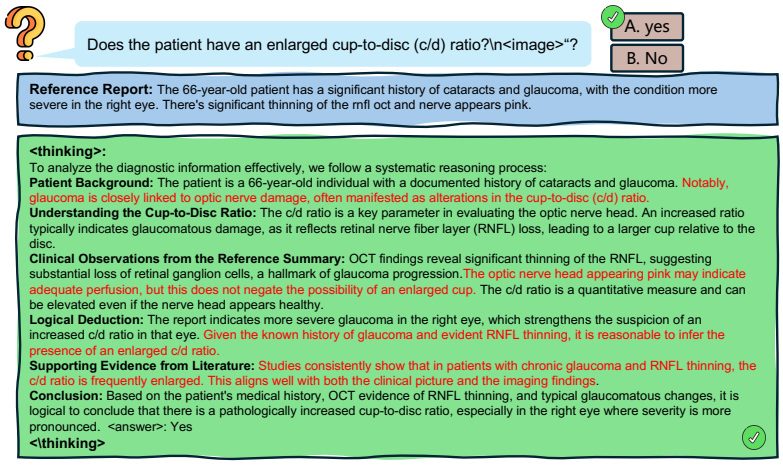

Fig. 2. Qualitative example of the generated chain-of-thought (CoT) reasoning.

Implementation Details. For the retriever, the vision encoder is a ResNet-50 [9], and the text encoder is a ViT-B/32 [5]. We use the AdamW optimizer with a learning rate of 10^{-3}, weight decay of 10^{-2}, and a batch size of 32. The model is trained for 360 epochs. For the multi-agent frame work, all of our agents are gpt-4omini and we used camel [16] and langgraph to build the entire workflow. For preference fintuning and the Med-LVLM, we use LLAVA-Med 1.5 7B [15] as the backbone model. During the preference optimization process, we adapt LoRA fine-tuning [10]. All experiments were performed using four NVIDIA A6000 GPUs on a high-performance computing (HPC) cluster.

For evaluation, we validated all three datasets across four fundamental metrics: Accuracy, Precision, Recall, and F1 Score to ensure a comprehensive performance assessment where Accuracy is considered as primary metric.

3.2 Comparison with State-of-the-arts

We compare our proposed method with three categories of recent state-of-the-art techniques: inference decoding, factual control, RAG-CoT, and the RAG-PT paradigm. For inference decoding, we evaluate two widely used methods: greedy decoding and beam search [27]. For factual control, we compare our approach against DoLa [3], OPERA [11], and VCD [13]. These methods are widely recognized as effective strategies for mitigating hallucinations in large language models. We include the results for these baselines from [34]. For RAG-CoT method, we have implemented RAT [29] in our current experimental setup. For RAG-PT paradigms, we evaluate our method against MMed-RAG [33] and RULE [34], both of which are specifically designed for medical multimodal RAG.

As shown in Table 1, our proposed model significantly outperforms other comparative methods on several validation metrics. For accuracy, GoCa improved 13.18% on average over the baseline and 4.91% on average over the

Table 1. Main Experimental results. The best results in each column are highlighted in **bold**, and the second-best values are underlined.

Models	IU-Xray				Harvard-FairVLM				MIMIC-CXR			
	Acc	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
LLaVA-Med-1.5	77.30	55.64	73.94	63.49	63.05	93.65	61.24	74.06	75.62	81.07	78.92	79.98
+ Greedy	79.90	58.58	74.40	65.55	62.38	93.98	60.16	73.36	58.90	86.54	61.74	72.07
+ Beam Search	75.47	52.64	81.07	63.83	68.54	**95.70**	66.44	78.43	60.14	86.91	63.08	73.10
+ RAT	75.36	81.38	77.68	79.49	77.92	56.78	72.48	63.68	62.75	**94.47**	60.27	73.59
+ DoLa	78.00	55.96	82.69	66.75	76.87	92.69	79.40	85.53	81.35	80.94	81.07	85.73
+ OPEAR	70.59	44.44	**100.0**	61.54	71.41	92.72	72.49	81.37	69.34	72.04	79.19	76.66
+ VCD	68.99	44.77	69.14	54.35	65.88	90.93	67.07	77.20	70.89	78.06	73.23	75.57
+ MMed-RAG	64.67	42.65	93.88	58.66	84.76	86.49	97.53	91.68	79.36	75.96	**97.18**	85.27
+ RULE	74.85	52.13	71.03	60.13	85.67	86.23	**99.18**	92.26	76.39	81.16	80.21	80.68
+ GoCa(Ours)	**81.18**	**60.24**	86.89	**71.56**	**88.55**	89.74	98.31	**93.83**	**84.87**	85.86	90.24	**88.00**

Table 2. Ablation study on Harvard-FairVLM dataset. The best results in each column are highlighted in **bold**, and the second-best values are underlined.

Models					Accuracy	Precision	Recall	F1 Score
GPT-4o†					22.45	96.00	11.52	20.57
R	**PT**	**C1**	**C2**	**K**				
					63.05	93.65	61.24	74.06
✓					71.13	**95.67**	69.62	80.59
✓	✓				84.66	86.44	97.47	81.63
✓	✓	✓			84.73	86.56	97.39	91.66
✓	✓	✓	✓		87.11	87.58	**99.07**	92.80
✓	✓	✓	✓	✓	**88.55**	89.74	98.31	**93.83**

† We observed that GPT4o tends to refuse to answer Med-VQA questions under the same prompt, So here we have only counted the cases that answered. The refusal rate was 83.73%.

second-best method. This proves our hypothesis for the original question that: *Explicit thinking distillation is effective in Med-LVLM reasoning from retrieval enhancements.* Meanwhile, Fig 2 also demonstrated the effectiveness of the method from a qualitative perspective.

3.3 Ablation Study

As shown in Table 2, we conducted a detailed ablation analysis on each component and level of the proposed methodology. Specifically, our baseline is the native LlaVA-Med. **R** represents the results after incorporating text retrieval via the retriever (Sec 2.1). **PT** denotes the the result under preference fine-tuning without CoTs. **C1** denotes the results with the addition of CoT generation (Sec 2.2.1), while **C2** includes the refinement process (Sec 2.2.2). Finally, **K** corresponds to replacing the fixed top-k selection with an adaptive k strategy (Sec 2.4). Notably, comparing **C1** and **C2**, we observe that when the CoT quality exhibits an apparent hierarchical disparity, model performance improves

effectively. This further validates our concern regarding potential reward hacking in binarized DPO and supports the necessity of hierarchical treatment in CoT refinement (Sec 2.3). Furthermore, to mitigate potential performance advantages from model scale, we evaluated GPT-4o for extra ablation. As shown in Table 2, GPT struggled with the medical task using the same prompt, highlighting the potential to distill reasoning from general to domain-specific models.

4 Conclusion

We introduced GoCa, a multi-modal RAG system designed to enhance explicit reasoning within the RAG-PT paradigm, making Med-LVLMs more process-driven, transparent, and trustworthy in medical decision-making. By integrating a self-evolving multi-agent framework, Chain-of-Thought preference optimization, and a Monte Carlo-inspired retrieval refinement mechanism, GoCa significantly improved factual grounding and mitigated hallucinations. Experimental results on three Med-VQA datasets demonstrated that GoCa achieved superior factual accuracy, outperforming recent state-of-the-art approaches. Our work established a new paradigm for integrating explicit reasoning with factual retrieval in Med-LVLMs, paving the way for clinically trustworthy LVLM-assisted decision-making systems.

Acknowledgments. This work was supported by JST GTIE GAP Fund Program, Grant Number GTIE2024_EX10, Japan.

Disclosure of Interests. The authors declare no competing interests.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Anthropic: The claude 3 model family: Opus, sonnet, haiku. In: online (2024), https://api.semanticscholar.org/CorpusID:268232499
3. Chuang, Y.S., Xie, Y., Luo, H., Kim, Y., Glass, J., He, P.: Dola: Decoding by contrasting layers improves factuality in large language models. arXiv preprint arXiv:2309.03883 (2023)
4. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., others J: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
6. Fan, L., Gong, X., Zheng, C., Tan, X., Li, J., Ou, Y.: Cycle-vqa: A cycle-consistent framework for robust medical visual question answering. Pattern Recognition **165**, 111609 (03 2025). https://doi.org/10.1016/j.patcog.2025.111609

7. Gai, X., Zhou, C., Liu, J., Feng, Y., Wu, J., Liu, Z.: Medthink: Explaining medical visual question answering via multimodal decision-making rationale. CoRR (2024)
8. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
11. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13418–13427 (2024)
12. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., et al.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
13. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13872–13882 (2024)
14. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
15. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
16. Li, G., Hammoud, H., Itani, H., Khizbullin, D., Ghanem, B.: Camel: Communicative agents for" mind" exploration of large language model society. Advances in Neural Information Processing Systems **36**, 51991–52008 (2023)
17. Li, J., Cao, P., Chen, Y., Liu, K., Zhao, J.: Towards faithful chain-of-thought: Large language models are bridging reasoners. arXiv preprint arXiv:2405.18915 (2024)
18. Li, J., Skinner, G., Yang, G., Quaranto, B.R., Schwaitzberg, S.D., Kim, P.C., Xiong, J.: Llava-surg: towards multimodal surgical assistant via structured surgical video learning. arXiv preprint arXiv:2408.07981 (2024)
19. Li, X., Mei, S., Liu, Z., Yan, Y., Wang, S., Yu, S., Zeng, Z., Chen, H., Yu, G., Liu, Z., et al.: Rag-ddr: Optimizing retrieval-augmented generation using differentiable data rewards. arXiv preprint arXiv:2410.13509 (2024)
20. Luo, Y., Shi, M., Khan, M.O., Afzal, M.M., Huang, H., Yuan, S., Tian, Y., Song, L., Kouhana, A., Elze, T., et al.: Fairclip: Harnessing fairness in vision-language learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12289–12301 (2024)
21. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H). pp. 353–367. PMLR (2023)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)

23. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems **36**, 53728–53741 (2023)
24. Shi, C., Rezai, R., Yang, J., Dou, Q., Li, X.: A survey on trustworthiness in foundation models for medical image analysis. arXiv preprint arXiv:2407.15851 (2024)
25. Skalse, J., Howe, N., Krasheninnikov, D., Krueger, D.: Defining and characterizing reward gaming. Advances in Neural Information Processing Systems **35**, 9460–9471 (2022)
26. Sun, L., Zhao, J., Han, M., Xiong, C.: Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. arXiv preprint arXiv:2407.15268 (2024)
27. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. Advances in neural information processing systems **27** (2014)
28. Suzuki, K.: Overview of deep learning in medical imaging. Radiological physics and technology **10**(3), 257–273 (2017)
29. Wang, Z., Liu, A., Lin, H., Li, J., Ma, X., Liang, Y.: Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. arXiv preprint arXiv:2403.05313 (2024)
30. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
31. Wu, T., Lan, J., Yuan, W., Jiao, J., Weston, J., Sukhbaatar, S.: Thinking llms: General instruction following with thought generation. arXiv preprint arXiv:2410.10630 (2024)
32. Xia, P., Chen, Z., Tian, J., Yangrui, G., Hou, R., Xu, Y., Wu, Z., Fan, Z., Zhou, Y., Zhu, K., et al.: CARES: A comprehensive benchmark of trustworthiness in medical vision language models. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024)
33. Xia, P., Zhu, K., Li, H., Wang, T., Shi, W., Wang, S., Zhang, L., Zou, J., Yao, H.: MMed-RAG: Versatile multimodal RAG system for medical vision language models. In: The Thirteenth International Conference on Learning Representations (2025)
34. Xia, P., Zhu, K., Li, H., Zhu, H., Li, Y., Li, G., Zhang, L., Yao, H.: Rule: Reliable multimodal rag for factuality in medical vision language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 1081–1093 (2024)
35. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)