MedThink: Inducing Medical Large-scale Visual Languange Models to Hallucinate Less by Thinking More

Anonymous ACL submission

Abstract

When Large Vision Language Models (LVLMs) are applied to multimodal medical generative tasks, they suffer from significant model hallucination issues. This severely impairs the model's generative accuracy, making it challenging for LVLMs to be implemented in real-world medical scenarios to assist doctors in diagnosis. Enhancing the training data for downstream medical generative tasks is an effective way to address model hallucination. Moreover, the limited availability of training data in the medical field and privacy concerns greatly hinder the model's accuracy and generalization capabilities. In this paper, we introduce a method that mimics human cognitive processes to construct fine-grained instruction pairs and apply the concept of chain-of-thought (CoT) from inference scenarios to training scenarios, thereby proposing a method called **MedThink**. Our experiments on various LVLMs demonstrate that our novel data construction method tailored for the medical domain significantly improves the model's performance in medical image report generation tasks and substantially mitigates the hallucinations. All resources of this work will be released soon.

1 Introduction

When Large Vision Language Models (LVLMs) [13, 7, 12] encounter complex and integrative medical problems, they require a nuanced understanding and reasoning capability akin to that of human doctors. The lack of robust reasoning capacities impedes these models from accurately diagnosing and assessing patient conditions. Consequently, when faced with complex integrative medical problems, LVLMs often produce seemingly accurate but fundamentally erroneous answers, known as hallucinations [2, 9], which hinder their practical application in real-world medical scenarios.

One approach to mitigating hallucinations is by increasing the quality as well as the diversity of the training data [5, 15] to increase the robustness and generalisation of the model, however, traditional data augmentation methods such as image flipping, rotation, scaling, and the introduction of synthetic noise are not suitable for the multimodal medical domain due to the high fidelity required in medical image interpretation. Any alteration that distorts the medical reality of the image or disrupts the feature alignment between images and text can lead to misdiagnoses or the overlooking of critical patient-specific details. Expanding data by rewriting text directly using large language models (LLMs) has received some favour [15, 8], but the increase in data diversity from it is extremely limited.

To address these challenges, we propose a hierarchical text enhancement method called **MedThink**, specifically designed for the medical domain. By mimicking human cognitive processes in analyzing a medical image, MedThink leverages the Large Language Model (LLM) to construct fine-grained instruction pairs and apply the concept of chain-of-thought (CoT) from inference scenarios to training scenarios. Through enhancing the hierarchy, depth, and diversity of medical text, MedThink encourages LVLMs to progressively observe and analyze from superficial appearances to deep pathological conditions and provide a reliable and comprehensive diagnosis. Experiments demonstrate that this approach significantly reduces models' medical hallucinations, overcoming the limitations of traditional hallucination-mitigating methods by data augmentation.

In summary, our contributions are as follows: **(i)** We introduce MedThink, a novel medical construction method that effectively mitigates hallucinations in LVLMs within the medical domain. **(ii)** We transform chaotic and unstructured medical reports into chain-like data for reasoning, creating a high-quality inferential IRG dataset. **(iii)** We demon-

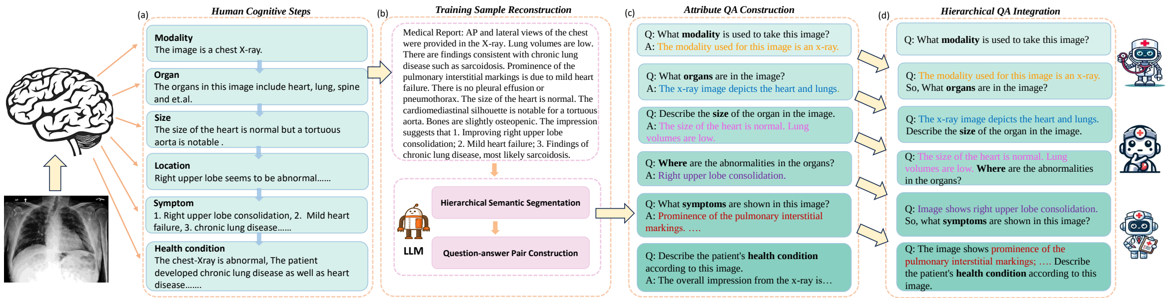


Figure 1: Illustration of MedThink's process for constructing hierarchical QA pairs based on real clinical image reports. (a). The progressive cognitive process of doctors for a medical image. (b). End-to-end data reconstruction of clinical reports. (c). Construction of Hierarchical QA pairs. (d). Chain-based QA Pair Refactoring

strate that applying MedThink to a baseline model achieves state-of-the-art (SOTA) performance in reducing medical hallucinations.

2 Methods

2.1 Overview of the MedThink

Medical texts authored by humans are often disorganized, complicating the process of correlating visual features with unstructured free texts. Therefore, we propose a hierarchical text enhancement method called **MedThink**, rooted in a chain-like hierarchy of medical attribute importance. This hierarchical semantic segmentation of medical text is inspired by the cognitive steps that doctors take when assessing medical images and diagnosing, involving the sequential recognition and integration of information across various levels.

Based on this cognitive process, we design prompts that enable powerful models like Gemini to generate structured question-answer pairs (QA pairs) end-to-end, focusing on different dimensions of medical images with fine granularity. Subsequently, we emulate the Chain-of-Thought (CoT) [3] paradigm and transfer this paradigm from the inference process into the model's fine-tuning process. This involves integrating lower-level informational cues into higher-level questions to generate new QA pairs and perform supervised fine-tuning (SFT) of the model.

2.2 Inspiration for the MedThink

Exploring whether human cognitive processes for image recognition can be applied to the organization of data in multimodal models is a question worth considering. Research in cognitive neuroscience indicates that humans typically process graphical information from superficial to deep levels [6], seamlessly understanding hierarchical information. However, AI models lack this capability and must analyze each image patch step by step by reducing the receptive field. Broadly speaking, the integration of textual information can assist AI models in emulating human cognitive processes.

Figure 1(a) illustrates the cognitive process of a doctor assessing a medical image. The process typically begins with global information about the image, such as the modality and the depicted human system, and progressively focuses on specific organs, followed by fine-grained attributes such as the shape or location of organs or abnormalities. After thoroughly understanding this information, the human brain synthesizes the symptoms presented in the image, ultimately producing an unstructured image report.

From the perspective of LVLMs, medical reports are usually highly disorganized. Even for humans, it is challenging to quickly correlate lengthy medical reports with every fine-grained semantic detail in the images. QA pairs, which provide a more granular depiction of dimensions, are often used as high-quality training data to help models understand images. However, constructing QA pairs is extremely time-consuming, and imaging centers typically produce image reports without explicitly creating QA pairs.

2.3 Application Details of Medthink

Construction of Hierarchical QA pairs: To transform disorganized image reports into high-quality QA pairs that mimic human cognition, we utilize powerful LLMs to semantically segment the image reports. As depicted in Figure 1(b), the image reports are processed by GPT-4 [1] for end-to-end semantic segmentation, and they are broken down into six dimensions: *modality*, *organ*, *size*, *abnormal location*, *symptoms*, and *overall health condition*. This segmentation results in six distinct questions, as shown in Figure 1(c), which independently query and answer the information from different dimensions of the image.

Chain-based QA Pair Refactoring: To further emulate human cognitive processes, we propose a chain-based QA pair refactoring method, illustrated in Figure 1(d). For each original question, the lower-level answer is used as a prelude to the higher-level question, thereby combining with the original question to reconstruct each QA pair in a chain-like manner.

2.4 Data Construction

Based on the aforementioned construction process in section 2.3, we obtained two sets of QA pairs: Hierarchical QA pairs and Chain-based QA pairs. The Chain-based QA pairs, combined with the original medical image reports, are used for the model's supervised fine-tuning (SFT). The Hierarchical QA pairs are treated as traditional Visual Question Answering (VQA) task data and can be used to evaluate the model's hallucination level (the images used for evaluation are different from those used for training).

Specifically, we collected 10,000 images with medical reports from the MIMIC dataset [10] and the OpenI [16] dataset. Using the MedThink method, we constructed 60,000 Hierarchical QA pairs and 60,000 Chain-based QA pairs. One-tenth of the medical image-text pairs from the OpenI dataset were used to test model performance, while the images from the MIMIC dataset were entirely used as test data to better evaluate the model's generalization capability.

2.5 Traing of baseline models

To demonstrate the superiority of our method, we conducted full-parameters SFT on several different types of baseline models, including LLaVA-Med(SOTA) [11], MiniGPT4 [18], XrayGPT [14], and mPLUG-Owl2 [17], and used the checkpoint provided in the original article.

2.6 Evaluation Metrics

We measured the Image Report Generation (IRG) capacities of the models with common metrics such as BERTScore, METEOR, ROUGE-1/2/L, and BLEU. Additionally, to verify the effectiveness of MedThink in mitigating model hallucinations by guiding Med-LVLMs to simulate human doctors' chain-like thoughts, we employed the novel medical text hallucination metric, the MediHall Score, which was proposed in recent advanced research [4]. This metric defines five common types of hallucinations specific to medical texts and ultimately calculates a comprehensive MediHall Score. The calculation formula for the MediHall score is given as:

$$\text{MediHall score} = \frac{1}{N}\sum_{i=1}^{N} S_i$$

where N is the total number of sentences in the report, and S_i is the score of the i-th sentence based on its hallucination type.

3 Experiment Results and Discussion

3.1 Is MedThink more effective than normal LLM data augmentation methods?

Table 1 presents a comparison of the performance of models fine-tuned using the MedThink method, models fine-tuned using the original IRG task data, and models fine-tuned using data simply expanded by GPT. The performance is evaluated on the IRG task test set. "R-1/2/L" means the ROUGE-1/2/L. "BS", "MT" stand for the BERTScore and METEOR, and MediHall means the MediHall Score.

As shown in Table 1, although the simple expansion of data using GPT significantly increased the volume of training data, it did not lead to performance improvements. This is reflected in a slight decrease in traditional metrics and a significant drop in the hallucination score, with a reduction of approximately 2% to 4% across the four models. In contrast, models trained using the MedThink method not only demonstrated improvements in traditional metrics but also surpassed the models trained with conventional data paradigms by approximately 2% to 5% in the hallucination score. This indicates that the MedThink method effectively enhances the model's understanding of the global information in images.

3.2 Are traditional data augmentation methods better at mitigating the illusion?

To further validate the enhancement of MedThink on the global understanding capability of models for medical images, we compared the MedThink method with traditional NLP data augmentation methods such as insert, swap, and delete, as well as

Table 1: Comparison results of the different augmentation methods on the OpenI dataset by various evaluation metrics. “R-1/2/L” means the ROUGE-1/2/L. “BS”, “MT” stand for the BERTScore and METEOR.

Model	BS	MT	R-1	R-2	R-L	BLEU	MediHall
LLaVA-Med + Origin	57.87	18.18	22.81	4.01	21.29	1.60	39.31
LLaVA-Med + GPT Rewriting	57.6	18.34	22.88	3.94	21.37	1.58	37.89
LLaVA-Med + MedThink	**65.67**	**28.44**	**32.5**	**8.73**	**29.99**	**5.82**	**41.36**
MiniGPT4 + Origin	60.18	19.82	24.59	5.12	25.12	1.51	62.03
MiniGPT4 + GPT Rewriting	59.41	20.10	24.06	4.97	21.30	1.50	60.31
MiniGPT4+MedThink	**61.32**	**21.37**	**25.14**	**5.97**	**26.10**	**1.61**	**65.62**
XrayGPT + Origin	64.71	20.97	30.24	6.97	27.58	1.69	64.11
XrayGPT + GPT Rewriting	64.69	21.30	30.22	6.59	27.65	1.60	61.33
XrayGPT + MedThink	**64.87**	**21.67**	**30.71**	**7.37**	**27.87**	**1.90**	**68.90**
mPLUG-Owl2+origin	69.77	35.38	40.39	12.91	36.67	7.81	68.80
mPLUG-Owl2+GPT rewriting	61.72	35.86	40.10	12.31	36.60	7.74	65.41
mPLUG-Owl2 + MedThink	**70.70**	**37.5**	**40.99**	**12.93**	**36.89**	**7.87**	**70.32**

Table 2: Comparison results of MedThink and traditional Data Augmentation methods on Medihall Score.

Model	MedThink	Insert	Swap	Delete	Blur	Flip	Clip
LLaVA-med	**41.36**	40.17	39.98	40.21	37.95	38.88	39.56
MiniGPT4	**65.62**	64.48	62.14	63.15	60.10	59.65	61.47
XrayGPT	**68.90**	66.55	62.81	55.27	55.97	598.6	62.01
mPLUG-Owl2	**70.32**	69.36	68.19	68.67	68.16	68.40	68.07

image augmentation methods like Gaussian noise, flip, and clip. The experimental results are shown in Table 2.

As observed from Table 2, while traditional data augmentation methods effectively expand the data volume and enhance the model’s robustness, they do not result in a significant improvement in the MediHall score. This further demonstrates that emulating human cognitive processes can directly reduce the hallucinations in LVLMs for medical applications, surpassing the noise-based traditional augmentation methods.

3.3 Do MedThink improve model generalisability?

To verify whether MedThink can enhance the model’s generalization capability, we evaluated the out-of-distribution (OOD) performance of models fine-tuned using chain-based QA pairs on images from the MIMIC dataset. The experimental results are presented in Table 3.

In Table 3, we compare the MedThink method with models fine-tuned using the original IRG task

Table 3: Out of Distribution (OOD) capacities of different models with different augmentation methods.

Model	BS	MT	R-1	R-2	R-L	BLEU	MediHall
LLaVA-Med	50.94	10.92	17.73	2.16	13.72	0.52	29.19
LLaVA-Med+origin	51.40	11.04	13.91	2.20	13.91	0.63	27.53
LLaVA-Med+MedThink	58.11	16.39	21.84	3.89	16.88	1.27	32.58
MiniGPT4	46.43	10.27	15.36	1.75	12.63	0.53	32.92
MiniGPT4+origin	60.54	23.09	25.55	4.84	19.94	1.84	41.11
MiniGPT4+MedThink	60.12	21.90	24.78	4.32	19.20	2.18	43.56
mPLUG-Owl2	55.01	09.91	18.21	2.14	13.01	0.12	35.03
mPLUG-Owl2+origin	55.16	09.60	18.10	2.15	13.23	0.14	36.63
mPLUG-Owl2+MedThink	55.48	10.01	18.42	2.19	13.49	0.15	38.58

Table 4: MediHall Score of LLaVA-Med and mPLUG-Owl2 on the VQA tasks.

	Origin	**MedThink**	Rewriting	Insert	Swap	Delete	blur	Flip	Clip
LLaVA-Med	61.29	**66.40**	60.33	62.15	60.58	60.93	61.45	60.46	60.95
mPLUG-Owl2	68.21	**73.98**	69.10	70.32	69.94	70.01	69.16	67.32	68.20

data and those fine-tuned using data simply expanded by GPT. XrayGPT, having been trained on the MIMIC dataset, is not applicable for the relevant performance evaluation. As shown in Table 3, models trained with the MedThink method significantly outperformed other methods in OOD performance. There were notable improvements in traditional metrics, and the hallucination score also improved by approximately 3% to 5%.

3.4 Can MedThink methods reduce hallucinations across the board?

To further substantiate the effectiveness of the MedThink method in mitigating model hallucinations and enhancing image understanding capabilities, we selected two models suitable for VQA tasks: LLaVA-Med and mPLUG-Owl2. We evaluated their performance using the Hierarchical QA pairs mentioned earlier. The experimental results are presented in Table 4.

Table 4 compares the hallucination scores of the untrained baseline models, models fine-tuned using the MedThink method, and models fine-tuned using traditional data augmentation methods. As shown in Table 4, traditional data augmentation methods did not alleviate the hallucination problem in the fine-grained VQA scenarios, with LLaVA-Med and mPLUG-Owl2 maintaining hallucination scores of 61.29% and 68.21%, respectively. In contrast, models trained with the MedThink method achieved a significant reduction in hallucination scores, demonstrating that MedThink effectively alleviates model hallucinations.

4 Conclusions

In this paper, we introduced MedThink, a hierarchical text enhancement method designed to address the hallucination issues faced by LVLMs in complex medical scenarios. MedThink leverages the cognitive processes of human doctors by segmenting medical text into different attribute layers and constructing chain-based QA pairs. Our experiments demonstrate that MedThink significantly enhances the performance of LVLMs in medical image report generation tasks, both in terms of traditional evaluation metrics and hallucination scores.

5 Limitations

Despite the promising results, there are several limitations to our study:

Evaluation Metrics: While the hallucination score and traditional metrics provide insights into model performance, developing more comprehensive evaluation metrics that capture the nuanced understanding required in medical diagnostics would be beneficial.

Model Complexity: The improvements brought by MedThink need to be assessed in the context of model complexity and inference time, as more sophisticated models may require more computational power and longer training times.

References

[1] 2023. Openai. GPT-4V(vision) system card. 2

[2] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*. 1

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165. 2

[4] Jiawei Chen, Dingkang Yang, et al. 2024. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.06805*. 3

[5] Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. 2023. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*. 1

[6] Alex Clarke and Lorraine K Tyler. 2015. Understanding what we see: how we derive meaning from vision. *Trends in cognitive sciences*, 19(11):677–687. 2

[7] W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*. 1

[8] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2024. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36. 1

[9] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38. 1

[10] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *Preprint*, arXiv:1901.07042. 3

[11] Chunyuan Li, Cliff Wong, Zhang, et al. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*. 3

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*. 1

[13] Haotian Liu, Chunyuan Li, Wu, et al. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*. 1

[14] Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*. 3

[15] Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer. 1

[16] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106. 3

[17] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*. 3

[18] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*. 3

A Appendix

A.1 Prompts for MedThink and Visualization

In the MedThink method, we use instruction as shown in Figure 2 to conduct semantic segmentation of the original medical reports, extract information at different levels, and then construct question-answer pairs respectively. We used ChatGPT to process the original medical report, which is divided into six question-answer pairs containing key information of different attributes, as shown in Figure 2. Then, as described in the Method section 2, the six pairs are sequentially combined into a chain structure containing progressive relationships.

In Figure 3, we show a comparison of the output results of mPLUG-Owl2 that performs best on the IRG task.

A.2 Experiments Compute Resources

In this paper, SFT for all baseline models was performed on 8 Nivida A800 GPUs, and Inference for all baseline models was performed on a single A800 GPU.

Medical Report Semantic Segmentation Instruction

Here is a medical report based on an image. Imagine you are an experienced doctor. Based on the report, please extract structured information of six attributes: image modality, organ, size, symptom, location, condition. Then, please construct six instructions based on the information of the six attributes. Here is an example you can refer to:

<medical report>: The x-ray image shows low lung volumes with bronchovascular crowding and no definite acute findings of focal alveolar consolidation or pleural effusion. However, a dense nodule in the right base suggests a previous granulomatous process. The heart size appears normal, and there are no typical findings of pulmonary edema. Overall, the x-ray impression is of low lung volumes with no acute abnormalities, but prior granulomatous disease is noted.

<six instructions>
1. <instruction>: What modality is used to take this image? <answer>: This is an x-ray.
2. <instruction>:What organs are in the image? <answer>: The lungs and heart are the main organs depicted in the x-ray image.
3. <instruction>: Please describe the size of the organ shown in the image. <answer>: The x-ray image shows low lung volumes, and the heart size appears normal.
4. <instruction>: What symptoms are shown in this image? <answer>: The x-ray image shows low lung volumes with bronchovascular crowding, and a dense nodule in the right base suggests a previous granulomatous process.
5. <instruction>: Where are the abnormalities (if any) in the organs? <answer>: The abnormality, a dense nodule, is located in the right base of the lungs.
6. <instruction>: Please describe the patient's health condition according to this image. <answer>: While there are no acute abnormalities, the presence of the granulomatous nodule indicates a prior pathological process affecting lung health.

Now, please provide six new instructions based on the following medical report:
<medical report>: {medical report}

Examples after Segmentation

[Original Report]
The findings from the X-ray image are normal, with no signs of active disease. The heart and lungs appear healthy, with no abnormalities detected. There are no signs of pleural effusion or pneumothorax, indicating that the patient's respiratory system is functioning properly. The PICC line is in place and the tip is located in the upper right atrium of the heart. Overall, the image suggests that the patient is in good health, with no immediate concerns

<instruction>: What modality is used to take this image?
<answer>: The modality used is X-ray imaging.

<instruction>: What organs are in the image?
<answer>: The heart and lungs are the main organs depicted in the X-ray image.

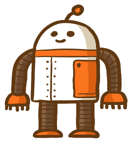
<instruction>: Please describe the size of the organ shown in the image.
<answer>: The report states that the heart and lungs appear healthy and normal, without specific size details provided.

<instruction>: What symptoms are shown in this image?
<answer>: The X-ray image indicates normal findings with no signs of active disease, pleural effusion, or pneumothorax, suggesting the absence of symptoms.

<instruction>: Where are the abnormalities (if any) in the organs?
<answer>: No abnormalities were detected in the heart or lungs. The PICC line is in place, with the tip located in the upper right atrium of the heart.

<instruction>: Please describe the patient's health condition according to this image.
<answer>: The image suggests that the patient is in good health, with no immediate concerns, and the respiratory system is functioning properly.

Figure 2: Medical Report Semantic Segmentation Instruction and the Example of Result



Question: Take a look at this chest x-ray and describe the findings and impression.

Ground-truth: The X-ray image shows no acute cardiopulmonary abnormality. The heart and lungs appear normal with no focal airspace disease, pneumothorax or pleural effusion. The cardiomedistinal silhouette and pulmonary vascularity are stable. The image indicates probable COPD and old granulomatous disease with mild biapical scarring. The lungs have low volumes but are clear bilaterally. CT scan may be warranted. No active disease is observed in the interval.

mPLUG-Owl2 & MedThink	The chest x-ray shows no signs of acute cardiopulmonary abnormality. Lungs have low volumes. The heart size and mediastinal contour are normal. There is no evidence of pneumothorax, pleural effusion or focal consolidation. The lungs are clear with no acute infiltrate or abnormalities. The impression is that there is no acute cardiopulmonary abnormality, but there may be granulomatous disease.
mPLUG-Owl2 & GPT Rewriting	According to the x-ray image, the heart size, cardiomediastinal silhouette, and pulmonary vasculature all appear to be within normal limits. Additionally, there are no infiltrates, effusions, or pneumothorax observed. Based on these findings, there is no evidence of an acute cardiopulmonary process present in the individual.
mPLUG-Owl2 & Insert	The x-ray shows that the cardiomediastinal silhouette is normal in size. The lungs are clear without any signs of pneumothorax or pleural effusion. The bony structures are normal. Overall, there are no acute bony abnormalities.
mPLUG-Owl2 & Delete	The chest x-ray shows normal lungs with no focal consolidation, pleural effusion, or pneumothorax. The heart is of normal size and the mediastinum is unremarkable. The cardia is mildly enlarged, but there are no acute cardiopulmonary abnormalities. The impression shows no acute cardiopulmonary abnormalities.
mPLUG-Owl2 & Swap	The x-ray shows clear lungs and a normal heart and mediastinal silhouette. The lungs are inflated with clear air spaces with no signs of any acute cardiopulmonary abnormality. The impression is that there is no acute cardiopulmonary abnormality present.
mPLUG-Owl2 & Gauss	The x-ray shows left midlung opacity with airspace consolidation and pneumothorax. The heart and mediastinal contours appear normal. There is no acute bony abnormality. This suggests that there are no active cardiopulmonary findings.
mPLUG-Owl2 & Flip	The x-ray shows clear lungs with a normal heart size. Although the heart and mediastinum are normal in size, there are findings of atelectasis and effusions. The impression suggests that there are some pleural effusions and pneumothoraces present. Further investigation is recommended.
mPLUG-Owl2 & Clip	Based on the X-ray image, the heart size and vascularity are normal, and the mediastinal contour is normal. The lungs are also clear, with no pleural effusions or pneumothoraces observed. Overall, the X-ray indicates a normal chest with no acute cardiopulmonary findings

Figure 3: Comparison of mPLUG-Owl2 outputs with different data enhancement methods