

HiA: Towards Chinese Multimodal LLMs for Comparative High-Resolution Joint Diagnosis

Xinpeng Ding^{1†}, Yongqiang Chu^{2†}, Renjie Pi¹, Hualiang Wang¹, and Xiaomeng Li^{1,3*}

¹The Hong Kong University of Science and Technology

² The Department of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology

³HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen

†Equal contribution *Corresponding author: cexmli@ust.hk

Abstract. Multimodal large language models (MLLMs) have been explored in the Chinese medical domain for comprehending complex healthcare. However, due to the flaws in training data and architecture design, current Chinese medical MLLMs suffer from several limitations: cultural biases from English machine translations, limited comparative ability from single image input and difficulty in identifying small lesions with low-resolution images. To address these problems, we first introduce a new instruction-following dataset, Chili-Joint (**C**hinese **I**nterleaved **I**mage-**T**ext Dataset for **J**oint Diagnosis) collected from the hospital in mainland China, avoiding cultural biases and errors caused by machine translation. Besides one single image input, Chili-Joint also has multiple images obtained at various intervals during a patient’s treatment, thus facilitating an evaluation of the treatment’s outcomes. We further propose a novel HiA (**H**igh-resolution instruction-aware **A**dapter) to incorporate high-resolution instruction-aware visual features into LLMs to facilitate the current MLLMs to observe the small lesions as well as the comparative analysis. Extensive experiments on Chili-Joint demonstrate our HiA can be a plug-and-play method to improve the performance of current MLLMs for medical analysis. The code is available at <https://github.com/xmed-lab/HiA>.

Keywords: Chinese multimodal data · Large language model · Adapter.

1 Introduction

The recent surge in Large Language Models (LLMs) development, exemplified by GPT4 [21], Vicuna [5] and LLama [25], has revolutionized language tasks through advanced algorithms and massive data sets. This innovation has extended into multimodal large language models (MLLMs) [13,18,31,7,2,6,14,10,15,22], integrating visual data to enhance language understanding. Some researchers have explored the potential to apply MLLMs to the medical domain. For example, LLaVA-Med [12] and MedVInT [32] enhance medical visual instruction by

leveraging PubMed Central’s captions [24], while Med-Flamingo [20] and Med-BLIP [3] innovate in medical visual question answering, extending to 3D data analysis like MRI. Med-PaLM M [26] and RadFM [29] streamline multiple tasks, enhancing medical AI’s scope and efficiency.

To address the limitation of English-centric models in Chinese medical applications, Qilin-Med-VL [19], the pioneering Chinese medical MLLM, was developed to analyze both textual and visual medical data. This model, refined through a two-phase training on extensive Chinese visual-text pairs, excels at producing medical captions and resolving intricate medical inquiries in Chinese. Nonetheless, Qilin-Med-VL and similar Chinese MLLMs face challenges due to training data and design issues. Firstly, the reliance on ChiMed-VL [19], a dataset translated from English by GPT-3.4, raises concerns about biases and errors introduced during translation, potentially compromising model reliability. Additionally, the current focus on single-image input restricts the models’ diagnostic capabilities, overlooking the clinical need for comparing multiple radiographic images to assess treatment outcomes effectively.

To handle the drawbacks of current Chinese medical multimodal datasets, we introduce a new dataset Chili-Joint (**C**hinese **I**nterleaved **I**mage-**T**ext Dataset for **J**oint Diagnosis). Chili-Joint has two important advantages. First, the image-text pairs in Chili-Joint are collected from one top tertiary hospital in mainland China, which avoids cultural biases and errors caused by data machine-translated from English. Second, our Chili-Joint has interleaved vision language context; all samples contain sequences of inter-related images and texts, *e.g.*, X-rays obtained at various intervals during a patient’s treatment and corresponding descriptions. This enables the comparison across different periods, thus facilitating an evaluation of the treatment’s outcomes.

By evaluating our proposed Chili-Joint, we find that current MLLMs (both medical and general domain) fail to detect small lesions and conduct the comparative analysis of images from different periods, which may be caused by two reasons. ❶ **Low-resolution input.** Due to the data, memory and computation cost limitations, prevalent methods invariably input images with a small size (*i.e.*, 336×336) for training [12,34]. However, the lesions often occupy only a small portion of the entire image, making it challenging to observe in low-resolution pictures [1], leading to diagnostic errors (see Fig. 1 (a)). ❷ **Instruction-agnostic visual extraction.** Current MLLMs use a pre-trained connector layer, *e.g.*, Q-Former [13], to extract and translate visual features into tokens recognizable by LLMs. However, this pre-trained lightweight layer is instruction-irrelevant and pre-trained on image-caption pairs, thus tending to the primary visual contents which are just enough for the captioning task, but omit other visual details, *i.e.*, changes in lesions across different periods (see Fig. 1 (b)).

To address the above problems, we propose a lightweight HiA (**H**igh-resolution instruction-aware **A**dapter) which can efficiently enable the current Chinese medical MLLMs to receive multiple high-resolution images and conduct a comparative analysis for evaluating the effect of therapy (see Fig. 1). Specifically, HiA consists of three components: a high-resolution visual encoder to produce high-

	CN: <image>此图是否有异常? EN: <image> Whether any abnormality in this image?
Qilin-Med-VL	CN: 可见关节诸骨骨质完整, 骨小梁显示清晰, 无明显异常。 EN: The visible bones of the joint are intact, and the trabecular bone is clearly visible without any obvious abnormalities.
Ours	CN: 右侧股骨下段可见片状低密度影。膝关节间隙未见变窄。 EN: A flake-like low-density shadow can be seen in the lower segment of the right femur. No narrowing of the knee joint space.
Ground truth	CN: 右侧股骨下段骺骨投影区可见片状低密度影。 EN: A flake-like area of low density can be seen in the projection area of the patella in the lower segment of the right femur.
(a) Small Lesions	
	CN: <image> 和 <image> 分别是患者第一次检查的图片和经过手术治疗后的图片, 请对比两张图片。 EN: <image> and <image> depict the patient's condition before and after surgical treatment, respectively. Please compare two images.
Qilin-Med-VL	CN: 第一张为左侧X光图像, 显示骨质边缘增生、硬化。第二张为左侧膝盖的X光图像, 显示骨质边缘增生、硬化。 EN: The first image is an X-ray of the left side, showing bone edge proliferation and hardening. The second hardening.
Ours	CN: 第一张显示左膝关节骨质边缘增生,硬化..... 第二张相比第一张, 增生, 硬化等情况减弱, 表明左侧膝关节术后改变。 EN: The first image shows proliferation and hardening of the bone edges at the left knee joint the second image shows a reduction in proliferation and hardening, indicating post-operative changes in the left knee joint.
Ground truth	CN: 第二张相比第一张, 左侧骨质边缘增生,硬化,髌间隆起变尖等程度减弱, 表明左侧膝关节术后改变。 EN: Compared to the first image, the second one shows a decreased extent of osteophyte formation, sclerosis, and sharpening of the intercondylar eminence on the left side, indicating post-surgical changes in the left knee joint.
(b) Comparative Diagnosis	

Fig. 1: **(a) Small lesions.** Due to the low-resolution input, small lesions (red box) can not be diagnosed by the current MLLM SOTA [19]. **(b) Comparative Diagnosis.** Since the instruction-agnostic visual extraction of connector layer, current methods fail to follow the instruction to conduct comparative diagnosis. Benefit from HiA, our method can accurately produce correct responses based on the instruction (see red words).

resolution visual features from high-resolution input, an instruction-aware extractor to capture instruction-related visual information from high-resolution features, and an injection module to inject the instruction-related visual information into LLMs for better understanding. Notably, our proposed HiA is training-efficient and plug-and-play to be applied to existing MLLMs. We freeze all parameters of current MLLMs and only fine-tune the HiA during the training stage. We conduct experiments on Chili-Joint and prove HiA can be a plug-and-play method to benefit current MLLMs for medical analysis.

2 Method

As shown in Fig. 2, our overall pipeline consists of two parts: a multimodal large language model (MLLM) (Section 2.1) and new proposed high-resolution instruction-aware adapter (HiA) (Section 2.2).

2.1 Existing Multimodal Large Language Model

Current medical or general MLLMs, *e.g.*, Qilin-Med-VL [19] or LLaVA [12], generally consist of four parts: a vision encoder, a connector, a tokenizer and a large language module, which are introduced briefly in the following. Note that we will show two images in each pair for illustration. More images can also be processed similarly.

Visual Encoder. The visual encoder of the MLLMs is generally the plain ViT initialized from CLIP [23], which has pre-trained on massive image-text pairs.

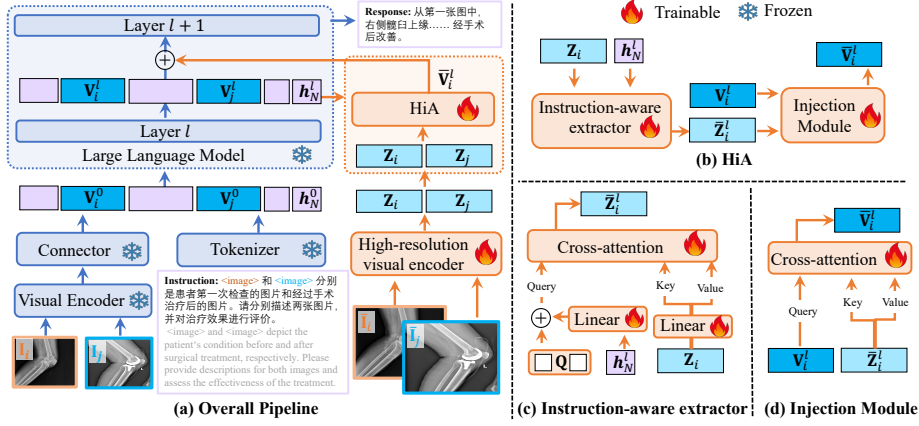


Fig. 2: Besides a general MLLM, which handles low-resolution image-text pairs (blue pathways), we also introduce a compact HiA mechanism (orange pathways) to improve MLLMs for comparative medical image analysis. The HiA module extracts visual information relevant to the instructions using an instruction-aware extractor and then incorporates this information into the MLLM through an injection module.

Formally, given the input images $\mathbf{I}_i$ and $\mathbf{I}_j$, the vision encoder produces the corresponding visual features $\mathbf{F}_i$ and $\mathbf{F}_j$ respectively.

Connector Layer. The lightweight connector layer has two purposes: (i) translate $\mathbf{F}_i$ into tokens $\mathbf{V}_i^0$ recognizable by LLMs and (ii) capture and compress the long visual features to fixed shorter ones. The connector is trained on millions of image-caption pairs by feeding the generated visual tokens, *i.e.*, $\mathbf{V}_i^0$, into a frozen LLM which generates the corresponding captions. Considering the computation and hardware cost, the size of the input image is in low-resolution, *e.g.*, generally up to 336×336 .

Tokenizer. The tokenizer aims to map the input text instructions to token embeddings, *i.e.*, $\mathbf{H}^0$, that are following fed into the LLM.

Large Language Model (LLM). As shown in Fig. 2 (a), before fed into the LLM, $\mathbf{V}_{i,j}^0$ and $\mathbf{H}^0$ are concatenated into a 1D sequence, formulated as follows:

$$\{\mathbf{h}_1^0, \mathbf{h}_2^0, \dots, \underbrace{\mathbf{v}_{i1}^0, \dots, \mathbf{v}_{ik}^0, \dots, \mathbf{v}_{iK}^0}_{\mathbf{V}_i^0}, \dots, \mathbf{h}_n^0, \dots, \underbrace{\mathbf{v}_{j1}^0, \dots, \mathbf{v}_{jK}^0}_{\mathbf{V}_j^0}, \dots, \mathbf{h}_N^0\}, \quad (1)$$

where $\mathbf{v}_{ik}^0$ is the k -th token in $\mathbf{V}_i^0$ and $\mathbf{h}_n^0$ is the n -th tokens in $\mathbf{H}^0$. In this paper, we follow Qilin-Med-VL [19] to use a renowned Chinese LLM, Chinese-LLaMA2-13B-Chat, as our pre-trained LLM.

2.2 High-Resolution Instruction-Aware Adapter

The high-resolution instruction-aware adapter (HiA) aims to capture high-resolution instruction-related visual information for MLLMs. Specifically, we first use a high-resolution visual encoder transfers the high-resolution images, *i.e.*,

$\bar{\mathbf{I}}_i$ and $\bar{\mathbf{I}}_j$, into visual features $\mathbf{Z}_i$ and $\mathbf{Z}_j$. The high-resolution visual encoder consists of several CNN [11] layers, which are more lightweight and training-efficient to handle dynamic resolutions of input, compared with the plain ViT [9] of MLLMs [4,8].

Given the high-resolution visual features and outputs from LLMs (including low-resolution visual tokens and textual tokens), HiA uses an instruction-aware extractor to capture instruction-related visual information from high-resolution features. Then, an injection module is adopted to incorporate the instruction-related visual information into LLMs for understanding and reasoning. We detail the instruction-aware extractor and the injection module in the following for $\mathbf{V}_i^l$, and so do as $\mathbf{V}_j^l$.

Instruction-Aware Extractor. We denote the output from the l -th layer of the LLM as $\{\mathbf{h}_1^l, \mathbf{h}_2^l, \dots, \mathbf{V}_i^l, \dots, \mathbf{h}_n^l, \dots, \mathbf{V}_j^l, \dots, \mathbf{h}_N^l\}$, where $\mathbf{V}_i^l = \{\mathbf{v}_{i1}^l, \dots, \mathbf{v}_{iK}^l\} \in \mathbb{R}^{K \times D_1}$ and $\mathbf{V}_j^l = \{\mathbf{v}_{j1}^l, \dots, \mathbf{v}_{jK}^l\} \in \mathbb{R}^{K \times D_1}$. As shown in Fig. 2 (c), we extract the last token $\mathbf{h}_N^l \in \mathbb{R}^{1 \times D_1}$ that can fully perceive the whole multimodal context during the first l layers and contains comprehensive instruction-aware semantics. Next, we obtain a set of learnable instruction-aware queries by:

$$\mathbf{Q} = \mathbf{Q} + \text{Linear}(\mathbf{h}_N^l), \quad \mathbf{Q} \in \mathbb{R}^{M \times D}, \quad \text{Linear}(\mathbf{h}_N^l) \in \mathbb{R}^{1 \times D}. \quad (2)$$

Finally, we use a cross-attention block, where $\mathbf{Q}$ as the query, high-resolution visual features $\mathbf{Z}_i$ as the value and key, to obtain $\bar{\mathbf{Z}}_i^l \in \mathbb{R}^{M \times D}$, which can be formulated as $\bar{\mathbf{Z}}_i^l = \text{CrossAtten}(\mathbf{Q}, \text{Linear}(\mathbf{Z}_i))$. In this way, $\bar{\mathbf{Z}}_i^l$ would contain more information related to the instructions. For example, given the instructions ‘assess the effectiveness of treatment’, the model would focus more on the lesion difference between two images to conduct an assessment, while the connector layer in existing MLLMs tends to capture the salient information.

Injection Module. After obtaining the instruction-related features $\bar{\mathbf{Z}}_i^l$, we use the injection module to interact the information between $\bar{\mathbf{Z}}_i^l$ and $\mathbf{V}_i^l$ by $\bar{\mathbf{V}}_i^l = \text{CrossAtten}(\mathbf{V}_i^l, \bar{\mathbf{Z}}_i^l)$, where $\bar{\mathbf{V}}_i^l \in \mathbb{R}^{K \times D_1}$. Finally, we feed the addition, *i.e.*, $\bar{\mathbf{V}}_i^l + \mathbf{V}_i^l$, into the $l + 1$ -th layer of the LLM. See Fig. 2 (d) for details.

Note that we only fine-tune the newly introduced high-resolution visual encoder and HiA, and keep other parameters of MLLMs frozen for data and training efficiency. Compared with previous MLLMs towards high-resolution input [8], our HiA considers challenges in medical images, *e.g.*, comparative, instruction-related extraction.

3 Experiments

Dataset. Prior work, such as Huatuo-26M [27] and Qilin-Med-VL [19], found that translating from English introduces biases and inaccuracies, compromising robustness. We collect our dataset, *i.e.*, Chili-Joint, from our collaborating hospitals for the native Chinese dataset. There are totally 14K interleaved image-text pairs are randomly split into training, validation and testing set at 7 : 1.5 : 1.5

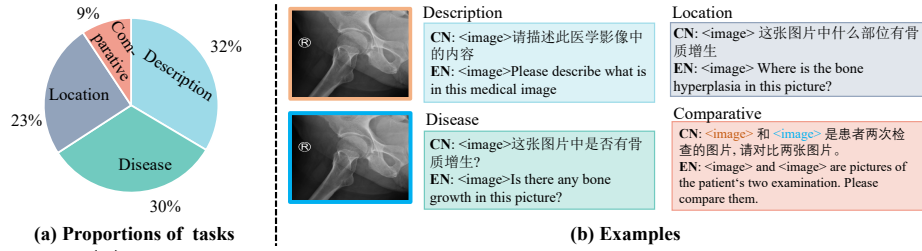


Fig. 3: (a) **Proportions of tasks.** The size of the arc represents the proportions of each task. (b) **Examples of different tasks.** For clarity, we only show instructions, omitting the responses.

ratio. We construction four different tasks, *i.e.*, Description, Disease, Location and Comparative, as shown in Fig. 3 (a). The detailed examples for each task are illustrated in Fig. 3 (b).

Implementation Details. We use Qilin-Med-VL [19]¹ as our baseline MLLM, which uses Chinese-LLaMA2-13B-Chat² as the foundation LLM and Clip-ViT-large-patch14-336 [23]³ as the pre-trained image encoder. Chinese-LLaMA2-13B-Chat is an open-source transformer-based LLM with 13 billion parameters further trained on Chinese-LLaMA2-13B and optimized for conversation scenarios. Clip-ViT-large-patch14-336 is a pre-trained CLIP vision encoder trained by OpenAI. The number of learnable instruction-aware queries is set to 256 (see Table 4a for ablation study). We select two layers of the LLM ($L/3$ and $2L/3$) to use our HiA (see Table 4b for analysis). As for the vision-language instruction-tuning stage, we use the following settings: batch size = 4 per GPU, one epoch, learning rate = $2e - 5$, warmup ratio = 0.03, and max length = 2048.

Evaluation Metrics. To evaluate the generated response based on the instructions, we use standard caption metrics [28,30,33,16], *i.e.*, BLEU-4 (B4), METEOR (M) to compare the consistency between the predictions from models and ground-truth.

3.1 Comparison with the State-of-the-Art Methods

To evaluate our proposed HiA, we select four state-of-the-art MLLMs including natural domain (*i.e.*, MiniGPT-4 [34] and LLaVA-1.5 [17]) and medical domain (*i.e.*, LLaVA-Med [12] and Qilin-Med-VL [17]). Among them, Qilin-Med-VL [17] is pre-trained on the Chinese medical multimodal dataset. For fair evaluation, all models are fine-tuned on the training set of Chili-Joint, select the model that achieving the best performance on the validation set, and report the performance on the test set on Table 1.

We observe two findings from Table 1. (i) Pre-training on medical data is essential, *i.e.*, models pre-trained on medical data outperform those not pre-trained. For example, the average performance of LLaVA-Med and Qilin-Med-VL across all tasks is 35.1 and 37.7, respectively, exceeding that of MiniGPT-4

¹ <https://github.com/williamliujl/Qilin-Med-VL>

² <https://github.com/LlamaFamily/Llama-Chinese>

³ <https://huggingface.co/openai/clip-vit-large-patch14-336>

Method	Description		Disease		Location		Comparative		AVG
	B4	M	B4	M	B4	M	B4	M	
MiniGPT-4 [34]	38.5	28.7	36.7	23.6	29.2	18.8	32.6	20.9	28.6
+ HiA	41.2	30.4	44.5	31.3	36.7	23.8	39.5	26.0	34.2
LLaVA-1.5 [17]	39.6	29.9	37.5	24.2	30.1	18.9	33.8	21.3	29.4
+ HiA	43.1	31.1	44.7	34.9	37.5	24.2	41.3	28.8	35.7
LLaVA-Med [12]	48.4	34.3	42.7	30.7	35.2	23.8	36.9	28.5	35.1
+ HiA	53.2	37.2	47.8	36.3	40.8	29.0	39.2	32.4	39.5
Qilin-Med-VL* [19]	51.4	38.5	44.6	33.2	38.7	26.8	37.8	30.7	37.7
+ HiA	54.8	40.6	49.5	38.6	43.9	28.0	42.5	33.5	41.4

Table 1: **Comparison with the state-of-the-art on Chili-Joint.** For all metrics, the higher the scores, the better the results. ‘B4’ and ‘M’ refer to BLEU-4 and METEOR. ‘AVG’ is the average value of all metrics on both two dataset. * indicates the model is pre-trained on Chinese medical multimodal data. Note that all models are fine-tuning on Chili-Joint in the same setting.

and LLaVA-1.5, which is 28.6 and 29.4. **(ii)** Our proposed HiA yields significant improvements for all models across all tasks. For instance, in terms of average performance, HiA achieves improvements of 5.6, 6.3, 4.5, and 3.7 compared to the four state-of-the-art models (MiniGPT-4, LLaVA-1.5, LLaVA-Med, and Qilin-Med-VL), respectively.

3.2 Ablation Study

In this section, we conduct the ablation study to evaluate the effectiveness of the proposed modules in HiA. We use Qilin-Med-VL [19] as our baseline model. **Effect of different proposed modules.** Our HiA model outperforms existing MLLMs by incorporating two novel types of information: instruction-related visual features and high-resolution (HR) features. To assess HiA’s impact, we use Qilin-Med-VL equipped with our HiA as a full model and conducted a series of experiments by systematically removing each component. Specifically, ‘w/o IA’ means that we remove the last token $\mathbf{h}_N^l$ (Eq. 2) from the learnable instruction-aware queries, thus ignoring the instruction semantics. ‘w/o HR’ indicates that we set the resolutions of high-resolution images to 336, equal to low-resolution images. The comparative analysis, summarized in Table 2, yielded two primary insights: **(i)** IA would not degrade the description task too much, since the original connector is trained to handle description-related tasks. IA module can capture more useful information; without IA, the performance of the other three tasks would degrade clearly. **(ii)** Without HR, the model fails to detect small lesions, resulting in descriptions that overlook these lesions and consequently degrade performance (also see Fig. 1 (a)).

Comparison of the baseline and ours across different resolutions. An intuitive way to enhance the perception ability is to increase the resolution of inputs, *i.e.*, inputs of varying resolutions are fed into the baseline and our method,

Method	Des	Dis	Loc	Com
Full	54.8	49.5	43.9	42.5
w/o IA	53.1	45.6	38.9	38.8
Δ	-1.7	-3.9	-5.0	-3.7
w/o HR	51.6	44.8	42.3	41.6
Δ	-3.2	-4.7	-1.6	-0.9

Table 2: **The ablation study of different proposed modules.** ‘Des’, ‘Dis’, ‘Loc’ and ‘Com’ indicate the Description, Disease, Location and Comparative tasks. ‘IA’ and ‘HR’ indicate using instruction-aware tokens and high-resolution input, respectively. Δ is the difference between SOTA with and without our HiA.

Res.	Memory	FLOPs	AVG
448	30.8 / 29.6	46.8 / 15.8	37.7 / 39.6
560	31.3 / 29.9	85.5 / 16.2	38.0 / 40.8
672	$\times$ / 30.1	105 / 16.5	$\times$ / 41.4
784	$\times$ / 30.3	162 / 16.7	$\times$ / 41.5
896	$\times$ / 30.4	208 / 16.8	$\times$ / 41.5

Table 3: **Comparison of the memory cost and flops between the baseline and our method.** $\times$ indicates the results cannot be obtained caused of out-of-memory. ‘Res.’ means the image resolution of high-resolution images. The results of the baseline model and our HiA are reported as red and blue respectively.

M	Des	Dis	Loc	Com
128	52.6	46.9	41.3	40.4
256	54.8	49.5	43.9	42.5
512	54.6	49.5	44.0	42.0

(a) M in Eq. 2 (Length of learnable instruction-related queries Q).

Num	Des	Dis	Loc	Com
1	52.1	45.8	41.2	39.9
2	54.8	49.5	43.9	42.5
3	54.2	49.0	43.4	41.8

(b) Number of layers of LLM incorporated with HiA.

Table 4: **Ablation on different designs in HiA.** Default settings are marked in gray.

respectively. To further study the effectiveness of our proposed HiA, we conduct experiments to compare the baseline and our method across different resolution inputs in Table 3. From the table, we can see that as the input image resolution increases, the memory cost and FLOPs of the baseline model grow proportionally, and would be out-of-memory when the input size rises to 672×672 . Differently, benefiting from HiA, our approach outperforms the baseline model while using much less computation and memory cost.

Effect of Different Designs. In our investigation detailed in Table 4, we examine the impact of varying design parameters: the length of learnable instruction-related queries (M) and the number of layers of LLM incorporated with HiA. Table 4a demonstrates that as the number of query tokens increases, the instruction-aware extractor would capture more useful information from high-resolution input, *e.g.*, $M = 256$ outperforms $M = 128$. However, too many query tokens, *e.g.*, 512, would bring more noise as well as computation cost, degrading the performance. Table 4b analyzes the effect of the different numbers of layers in the LLM to be injected with HiA. Specifically, Num = 1, Num = 2 and Num = 3 mean that we incorporated HiA in layer $\{L/2\}$, $\{L/3, 2L/3\}$ and $\{L/4, 2L/4, 3L/4\}$ respectively, where L is the total layer number of the LLM. Results show that incorporating HiA into two layers achieves the best performance.

4 Conclusion

In this paper, we propose a new dataset Chili-Joint (**C**hinese **I**nterleaved **I**mage-**T**ext Dataset for **J**oint Diagnosis), featuring image-text pairs from a leading tertiary hospital in China to mitigate cultural biases and enable comparative analysis of treatment through interrelated images and descriptions. We introduce HiA, a lightweight adapter that enhances Chinese medical MLLMs’ ability to analyze multiple high-resolution images for therapy evaluation. HiA, designed as a plug-and-play, training-efficient component, significantly improves medical MLLMs for medical analysis. Our current dataset is limited to X-ray images, and we plan to expand it by including more diverse modalities like CT and MRI in the future.

Acknowledgements. This work was supported in part by grants from the National Natural Science Foundation of China under Grant No. 62306254, grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Reference Number: T45-401/22-N), and Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYB-2020083).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baumgartner, M., Jäger, P.F., Isensee, F., Maier-Hein, K.H.: nndetection: a self-configuring method for medical object detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 530–539. Springer (2021)
2. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm’s referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
3. Chen, Q., Hu, X., Wang, Z., Hong, Y.: Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. arXiv preprint arXiv:2305.10799 (2023)
4. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions (May 2022)
5. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023) (2023)
6. Dai, W., Li, J., Li, D., Huat, A., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
7. Ding, X., Han, J., Xu, H., Liang, X., Zhang, W., Li, X.: Holistic autonomous driving understanding by bird’s-eye-view injected multi-modal large models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13668–13677 (2024)

8. Ding, X., Han, J., Xu, H., Zhang, W., Li, X.: Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. arXiv preprint arXiv:2309.05186 (2023)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2016). <https://doi.org/10.1109/cvpr.2016.90>, <http://dx.doi.org/10.1109/cvpr.2016.90>
12. 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** (2024)
13. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
14. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: *International Conference on Machine Learning*. pp. 12888–12900. PMLR (2022)
15. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
16. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 13753–13762 (2021)
17. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
19. Liu, J., Wang, Z., Ye, Q., Chong, D., Zhou, P., Hua, Y.: Qilin-med-vl: Towards chinese large vision-language model for general healthcare. arXiv preprint arXiv:2310.17956 (2023)
20. 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)
21. OpenAI, O.: Gpt-4 technical report (Mar 2023)
22. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PMLR (2021)
24. Roberts, R.J.: Pubmed central: The genbank of the published literature. *Proceedings of the National Academy of Sciences* p. 381–382 (Jan 2001). <https://doi.org/10.1073/pnas.98.2.381>, <http://dx.doi.org/10.1073/pnas.98.2.381>

25. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
26. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical ai. *NEJM AI* **1**(3), A10a2300138 (2024)
27. Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., Liu, T.: Huatuo: Tuning llama model with chinese medical knowledge. arXiv preprint arXiv:2304.06975 (2023)
28. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology* **1**(3), 100033 (2023)
29. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology. arXiv preprint arXiv:2308.02463 (2023)
30. Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G.R., Huang, X.: Multimodal recurrent model with attention for automated radiology report generation. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part I*. pp. 457–466. Springer (2018)
31. Zhang, H., Li, X., Bing, L., et al.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
32. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)
33. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 34, pp. 12910–12917 (2020)
34. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)