

Appendix

1 Templates of instruction-format data

We constructed an instruction-format medical image-text dataset to investigate their impact on fundamental VLMs. In detail, we transformed the QA pairs into an instruction fine-tuning format by constructing instruction templates, as shown in Figures 1 to 2. 'Closed' and 'Opened' templates are designed for closed-ended and open-ended questions respectively. Each QA pair is randomly embedded with one of these templates during training. To generate answer options, we classified the question attributes based on modality, plane, shape, size, organ, location, and pathology. Consequently, we created pools of candidate answers for different question attributes. Incorrect answers from the same attribute are randomly selected and embedded into the option together with the ground-truth answer.



Type	Templates
 Closed	1. <image> Question: {Question} Option: {Yes or No} The answer to the question is: 2. <image> Question: {Question} Option: {Yes or No} Answer: 3. <image> Question: {Question} Based on the image, give a judgmental answer to the question:
 Opened	1. <image> Question: {Question} Option: {A or B} The answer to the question is: 2. <image> Question: {Question} Option: {A or B} Answer: 3. <image> Question: {Question} Option: {A or B or C} Short answer: 4. <image> Used the provided image to answer the question: {Question} Option: {A or B or C} Provide your answer based on the option: 5. <image> What is the answer to the following question? "{Question}":

Fig. 1: Instruction-format Data Templates.



Original QA Pairs 	Instruction-format Data 	GT
Question: "Does the picture contain the organ which has the effect of discharging waste?" Answer: "No"	<image> Question: {Does the picture contain the organ which has the effect of discharging waste?} Option: {(a)No; (b)Yes} Short answer:	No
Question: "Which type of modality is shown about this image?" Answer: "CT"	<image> Used the provided image to answer the question: {Which type of modality is shown about this image?} Option: {(a)X-Ray; (b)MRI; (c) CT} Provide your answer based on the option	CT
Question: "What color does the spinal cord show in the picture?" Answer: "Gray"	<image> What is the answer to the following question? {What color does the spinal cord show in the picture?}	Gray

Fig. 2: Instruction-format Data Examples.

2 Details of Experiments

More Detailed Results of MILE Variants: As mentioned in Section 4 of our paper, we trained a series of MILE variants using the instruction-format data, and Tables 1 to 2 present the related results of four MILE variants on the Slake dataset. To more fully demonstrate the impact of different PEFT methods on the basic visual language model, we have also verified the PEFT method's impacts on BiomedGPT-Tiny: only the decoder is fine-tuned with PEFT methods, and the rest is full parameters fine-tuned. Table 3 presents the ACC on the benchmark, which shows a similar trend as MILE.

Implement Details: Here, we will present the experimental details of our model. All the training was conducted on a single NVIDIA RTX8000-48GB GPU. We used the Adamw optimizer with cosine learning rate decay, an initial learning rate of $2e-5$, a weight decay of 0.05, and a minimum learning rate of 0. The input image size for our model was 480×480 pixels, trained for 130 epochs, with a batch size of 20. For our baseline model, the ViT-based visual encoder consists of 12 layers of transformer, and both the JTM encoder and text decoder contain 12 layers of transformer-based layers.

Why choose this baseline model: We chose MISS as our baseline model and developed MILE based on it because MISS is a small-scale generative VLM and it has similar architecture to current LVLMs such as LLaVA, BLIP2, and et.al. MISS unifies the Text encoder and the Multimodal encoder by a JTM encoder so we can more easily fine-tune it.

Table 1: Results of MILE-LoRA (instruction-format data).

ViT	JTM	Dec	Rank	#Params	Memory	Opened	Closed	Gobal
F	LoRA	LoRA	4	0.163%	5.44	0	50.7	16.98
		LoRA	8	0.325%	5.53	0	50.7	16.98
LoRA	LoRA	LoRA	4	0.327%	26.90	28.09	29.01	28.40
			8	0.652%	27.01	48.93	24.82	31.51
F	T	LoRA	4	38.022%	7.85	21.79	35.77	26.49
		LoRA	8	38.072%	7.94	27.35	39.44	31.32
T	LoRA	LoRA	4	24.009%	26.95	39.57	8.73	29.25
		LoRA	8	24.133%	27.62	41.42	23.10	35.28
T	T	LoRA	4	61.887%	27.66	61.28	36.34	52.92
			8	61.919%	28.63	63.54	45.35	57.45

Table 2: Results of MILE-Prefix & IA3 & PTV2 (instruction-format data).

ViT	JTM	Dec	#Params	Memory	Opened	Closed	Gobal
F	IA3	IA3	0.051%	6.41	0	50.70	16.98
IA3	IA3	IA3	0.061%	23.47	0	50.70	16.98
T	IA3	IA3	23.924%	27.42	12.77	27.04	17.92
F	T	IA3	37.987%	8.18	8.37	27.04	14.62
T	T	IA3	61.866%	28.81	50.21	49.86	50.09
F	F	Prefix	3.926%	4.76	0	50.70	17.30
F	Prefix	Prefix	7.556%	4.81	0	50.70	17.30
T	Prefix	Prefix	29.636%	26.56	7.23	22.38	12.64
T	T	Prefix	63.354%	28.14	68.65	32.39	56.51
F	F	PTV2	0.051%	4.66	7.10	0	4.72
F	PTV2	PTV2	0.102%	4.70	0	0	0
T	PTV2	PTV2	23.963%	26.03	6.10	23.38	11.89
T	T	PTV2	61.876%	27.83	3.12	30.99	12.43

Table 3: Results on BiomedGPT-Tiny (origin data).

Method	#Params	Opened	Closed	Global
Full Fine-tuning	100%	71.84	64.46	68.97
Decoder-LoRA	50.76%	66.82	63.48	65.52
Decoder-Prefix	51.05%	69.94	60.54	66.29
Decoder-IA3	50.49%	64.95	52.21	60.01
Decoder-PTV2	50.92%	68.07	48.78	60.57