Towards All-in-One Medical Image Re-Identification Supplementary Material

Yuan Tian[1] Kaiyuan Ji[2] Rongzhao Zhang[1] Yankai Jiang[1] Chunyi Li[3]
Xiaosong Wang[1]✉ Guangtao Zhai[3]✉
[1]Shanghai AI Laboratory
[2]School of Communication and Electronic Engineering, East China Normal University
[3]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong Unversity
tianyuan168326@outlook.com

Model	Chest-X		Mess2		HCC-TACE	
	Acc↑	FP↓	Acc↑	FP↓	Acc↑	FP↓
InstructBlip [6]	46.40	100.00	50.00	100.00	50.00	100.00
LLaVA1.6 [9]	50.00	100.00	50.00	100.00	50.00	100.00
InternVL2 [4]	49.70	81.20	49.29	61.82	43.48	43.48
mPLUG-Owl3 [13]	50.10	99.60	47.58	90.31	50.00	0.00
Med-Flamingo [11]	50.30	100.00	50.71	100.00	50.00	100.00
HuaTuo [12]	70.00	45.00	57.12	84.33	54.35	17.39
QWen-VL-Max [3]	76.80	19.80	85.19	21.65	54.35	0.00
GPT-4o [2]	62.50	1.20	73.50	0.85	54.35	0.00
Ours	**98.80**	**2.23**	**93.82**	**3.76**	**95.27**	**2.85**

Table 1. Comparison of recent large multi-modal language models on the MedReID task. Acc and FP denotes Accuracy (%) and False positive rate (%), respectively. The best and the second best results are marked with **gray bold** and gray, respectively.

1. Comparison of Large Multi-modality Models.

We assess the performance of Large Multi-modality Models (LMMs) on ReID task, by using Yes/No question-answer pairs. We input the LMMs with a pair of medical images, and query if they are from the same patient.

Table 1 shows that general LMMs, such as LLaVA1.6 and mPLUG3, perform inadequately on this task, achieving only 50% accuracy, equivalent to random guessing. Among medical LMMs, the outdated Med-Flamingo underperforms across all modalities, while the more recent HuaTuo model demonstrates a much better ability to identify X-ray images from the Chest-X dataset. We also evaluate the current leading commercial LMMs, QWen-VL-Max and GPT-4o. These models achieve satisfactory results on single-image modalities like X-ray and Fundus. For example, QWen-VL-Max achieves 76.80% and 85.19% accuracy on Chest-X and Mess2. However, they struggle with multi-slice CT scans from HCC-TACE, indicating that even the most advanced LMMs are deficient in handling the MedReID task. In contrast, our model achieves high performance across all modalities, e.g., 98.80%/95.27% on Chest-X/HCC-TACE.

2. Training/Internal Validation Split

The Medical Image Re-Identification (MedReID) task is relatively new, and there is no established data splitting protocol. Using a fixed train/test set ratio such as 3:2 is inappropriate. For instance, in a very large dataset, the reserved number of test IDs can be substantial, posing significant challenges to algorithms and making it difficult to establish a meaningful benchmark. Conversely, in a smaller dataset, the reserved number of test IDs may be too few, leading to performance saturation for most methods.

Therefore, we adopt the following data splitting guidelines:

- For large-scale datasets such as MIMIC-X [8], we reserve 1000 IDs for the test set, while the remaining IDs are used for training.
- For other small-scale datasets such as CCII [14], we split the train/test ID ratio as 3:2.

The detailed splits are provided in Table 2.

Table 2. Train/Internal Val split protocol. 'Ab' denotes 'Abdominal'. For *EyePACS*, we do not reserve internal validation samples, since there are numerous fundus datasets for external evaluation.

Dataset	**Modality**	**Train/Val (IDs)**	**Train/Val (Images)**
MIMIC-X [8]	Chest X-ray	27852/1000	106333/5000
CCII [14]	Lung CT	601/401	1473/987
HCC-TACE [10]	Ab-CT	63/42	127/84
EyePACS [7]	Eye Fundus	17563/0	35126/0
ODIR [1]	Eye Fundus	1820/1214	3640/2428
LUAD [5]	Histopathology	75/50	325/127

3. Details for Large Multi-Modality Models (LMMs)

Data preprocessing and prompt configuration of different LMMs, on different datasets, are provided as follows.

3.1. Chest-X Dataset

This dataset was sampled from the original Chest-X dataset, containing 1,000 pairs of Chest X-ray images. Among these, 500 pairs were obtained from the same individual, while the remaining 500 pairs were sourced from different individuals.

InstructBlip: The model is the InstructBLIP-Vicuna-7B version. Two X-ray images are input into the model along with the following prompt.

> Can you please determine if these two X-ray images are from the same person? Only answer yes or no.

LLaVA1.6: This model is llava-v1.6-vicuna-7b. The two images are first resized to 256×256 and processed with the following prompt.

> Can you please determine if these two X-ray images are from the same person? Only answer yes or no, and provide reasons. Do not refuse judgment. The answer format is: yes/no.

InternVL2: This model is InternVL2-8B. During input processing, the images are resized to suitable dimensions and cropped into multiple 448×448 blocks. Additionally, a 448×448 thumbnail is generated and included in the list of image blocks. The, the processed images are then used as model input with the following prompt.

> Can you please determine if these two X-ray images are from the same person? Only answer yes or no, and provide reasons. Do not refuse judgment. The answer format is: yes/no.

mPLUG-Owl3: The model is mPLUG-Owl3-7B. The two X-ray images are input into the model with the following prompt.

> Can you please determine if these two X-ray images are from the same person? Only answer yes or no, and provide reasons. Do not refuse judgment. The answer format is: yes/no.

Med-Flamingo: The model is Med-Flamingo-9B. The two images are resized to 256×256 and input into the model with the following prompt.

> Can you please determine if these two X-ray images are from the same person? Do not refuse to make a judgment. [yes/no]

HuaTuo: The model is HuatuoGPT-Vision-7B. The two X-ray images are input into the model with the following prompt.

> Can you please determine if these two X-ray images are from the same person? Only answer "yes" or "no", and provide reasons. Do not refuse judgment. The answer format is: "yes" or "no".

QWen-VL-Max: This is a commercial model. During usage, the local paths of the two X-ray images are input into the model. The model automatically processes the images, sends requests via an API, and retrieves the corresponding responses. The prompt used is.

> Can you please determine if these two X-ray images are from the same person? Only answer "yes" or "no", and provide reasons. Do not refuse judgment. The answer format is: "yes" or "no".

GPT-4o: This is a commercial model. During usage, the two images are converted into base64 encoding and sent along with the prompt via the GPT-4o API to obtain a response. The prompt is.

> Can you please determine if these two X-ray images are from the same person? Only answer "yes" or "no", and provide reasons. Do not refuse judgment. The answer format is: yes/no.

3.2. Mess2 Dataset

This dataset is a subset from the original Mess2 dataset, containing 702 pairs of retinal fundus images. Among these, 351 pairs are from the same individual, while the remaining 351 pairs are from different individuals.

InstructBlip: The model is InstructBLIP-Vicuna-7B. During usage, the two fundus images are input into the model along with the following prompt.

> Can you determine whether these two images of retinal fundus are from the same person? Only answer "yes" or "no", and provide reasons. Do not refuse to make a judgment. The answer format is: "yes" or "no".

LLaVA1.6: This model is llava-v1.6-vicuna-7b. When processing, the two fundus images need to be resized to 256×256 and input into the model with the following prompt.

> Can you determine whether these two images of retinal fundus are from the same person? Only answer "yes" or "no", and provide reasons. Do not refuse to make a judgment. The answer format is: "yes" or "no".

InternVL2: This model is InternVL2-8B. During input processing, the images are resized to suitable dimensions and cropped into multiple 448×448 blocks. Additionally, a 448×448 thumbnail is generated and added to the list of image blocks. Then, the processed images are input into the model with the following prompt.

> Can you determine whether these two images of retinal fundus are from the same person? Only answer "yes" or "no", and provide reasons. Do not refuse to make a judgment. The answer format is: "yes" or "no".

mPLUG-Owl3: The model is mPLUG-Owl3-7B. During usage, the two fundus images are input into the model along with the following prompt.

> Can you determine whether these two images of retinal fundus are from the same person? Only answer "yes" or "no", and provide reasons. Do not refuse to make a judgment. The answer format is: "yes" or "no".

Med-Flamingo: The model is Med-Flamingo-9B. During usage, the two fundus images are resized to 256×256 and input into the model with the following prompt.

> Can you determine whether these two images of retinal fundus are from the same person? Do not refuse to make a judgment. [yes/no].

HuaTuo: The model is HuatuoGPT-Vision-7B. During usage, the two fundus images are input into the model with the following prompt.

> Can you determine whether these two images of retinal fundus are from the same person? Only answer "yes" or "no", and provide reasons. Do not refuse to make a judgment. The answer format is: "yes" or "no".

QWen-VL-Max: This is a commercial model. During usage, the local paths of the two retinal fundus images are directly input into the model. The model automatically processes the images, sends requests via an API, and retrieves the corresponding responses. The prompt used is.

> Can you determine whether these two images of retinal fundus are from the same person? Only answer "yes" or "no", and provide reasons. Do not refuse to make a judgment. The answer format is: "yes" or "no".

GPT-4o: This is a commercial model. During usage, the two retinal fundus images are converted into base64 encoding and sent along with the prompt via the GPT-4o API to obtain a response. The prompt is.

> Can you determine whether these two images of retinal fundus are from the same person? Only answer "yes" or "no", and provide reasons. Do not refuse to make a judgment. The answer format is: "yes" or "no".

3.3. HCC-TACE Dataset

This dataset is a subset from the original HCC-TACE dataset. The subset contains 46 pairs of CT modality images. Among them, 23 pairs of CT sequences are from the same individual, while the other 23 pairs are from different individuals. Considering the complexity of submitting entire CT sequences to large language models for identification, this study selected the first, middle, and last images from each CT sequence for analysis.

InstructBlip: The model used is InstructBLIP-Vicuna-7B. During usage, the CT sequence images are input into the model along with the following prompt.

> Can you determine whether the two CT sequences come from the same person? The first three images are from one sequence, and the last three are from another. Please answer with either "yes" or "no" and provide a reason. Do not refuse to make a judgment. The answer format is: "yes" or "no".

LLaVA1.6: This model is llava-v1.6-vicuna-7b. As the model cannot process the CT sequence images simultaneously, only the middle images from each CT sequence are used. The images are resized to 256×256 and input into the model with the following prompt.

> Can you determine whether the two CT images come from the same person? Only answer "yes" or "no", and provide reasons. Do not refuse to make a judgment. The answer format is: "yes" or "no".

InternVL2: This model is InternVL2-8B. During input processing, the CT sequence images are input into the model with the following prompt.

> Can you determine whether the two CT sequences come from the same person? The first three images are from one sequence, and the last three are from another. Please answer with either "yes" or "no" and provide a reason. Do not refuse to make a judgment. The answer format is: "yes" or "no".

mPLUG-Owl3: The model is mPLUG-Owl3-7B. During usage, the CT sequence images are directly input into

the model with the following prompt.

> Can you determine whether the two CT sequences come from the same person? The first three images are from one sequence, and the last three are from another. Please answer with either "yes" or "no" and provide a reason. Do not refuse to make a judgment. The answer format is: "yes" or "no".

Med-Flamingo: The model is Med-Flamingo-9B. During usage, the CT sequence images are resized to 256×256 and input into the model with the following prompt.

> Can you determine whether the two CT sequences come from the same person? The first three images are from one sequence, and the last three are from another. Do not refuse to make a judgment. [yes/no].

HuaTuo: The model is HuatuoGPT-Vision-7B. During usage, the CT sequence images are directly input into the model with the following prompt.

> Can you determine whether the two CT sequences come from the same person? The first three images are from one sequence, and the last three are from another. Please answer with either "yes" or "no" and provide a reason. Do not refuse to make a judgment. The answer format is: "yes" or "no".

QWen-VL-Max: This is a commercial model. During usage, the local paths of the CT sequence images are directly input into the model. The model automatically processes the images, sends requests via an API, and retrieves the corresponding responses. The prompt used is.

> Can you determine whether the two CT sequences come from the same person? The first three images are from one sequence, and the last three are from another. Please answer with either "yes" or "no" and provide a reason. Do not refuse to make a judgment. The answer format is: "yes" or "no".

GPT-4o: This is a commercial model. During usage, the CT sequence images are converted into base64 encoding and sent along with the prompt via the GPT-4o API to obtain a response. The prompt is.

> Can you determine whether the two CT sequences come from the same person? The first three images are from one sequence, and the last three are from another. Please answer with either "yes" or "no" and provide a reason. Do not refuse to make a judgment. The answer format is: "yes" or "no".

References

[1] Peking university international competition on ocular disease intelligent recognition (odir-2019). https://odir2019.grandchallenge.org/. Accessed: 2022-02-10. 1

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 1

[4] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 1

[5] National Cancer Institute Clinical Proteomic Tumor Analysis Consortium et al. The clinical proteomic tumor analysis consortium lung adenocarcinoma collection (cptac-luad)(version 12). the cancer imaging archive website. 1

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv*, 2023. 1

[7] Emma Dugas, Jared, Jorge, and Will Cukierski. Diabetic retinopathy detection. https://kaggle.com/competitions/diabetic-retinopathy-detection, 2015. Kaggle. 1

[8] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 1

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[10] A Moawad, D Fuentes, A Morshid, A Khalaf, M Elmohr, A Abusaif, JD Hazle, AO Kaseb, M Hassan, A Mahvash, et al. Multimodality annotated hcc cases with and without advanced imaging segmentation [data set]. *The Cancer Imaging Archive*, 2021. 1

[11] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 1

[12] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023. 1

[13] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou.

mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 1

[14] Kang Zhang, Xiaohong Liu, Jun Shen, Zhihuan Li, Ye Sang, Xingwang Wu, Yunfei Zha, Wenhua Liang, Chengdi Wang, Ke Wang, et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell*, 181 (6):1423–1433, 2020. 1