

Towards Scalable and General Multi-Modal Medical Data Analysis

@VALSE Webinar, September-18th-2024

Xiang Li (xli60@MGH.Harvard.edu)
[xiangli-shaun.github.io](https://github.com/xiangli-shaun)

- ❖ Evolution of “Alignment”
- ❖ Generative modeling: advantages and pitfalls
- ❖ BiomedGPT: A “foundation” approach

❖ Evolution of “Alignment”

Deep Learning-Based Image Segmentation on Multimodal Medical Imaging

Zhe Guo¹, Xiang Li¹, Heng Huang, Ning Guo, and Quanzheng Li¹

Abstract—Multimodality medical imaging techniques have been increasingly applied in clinical practice and research studies. Corresponding multimodal image analysis and ensemble learning schemes have seen rapid growth and bring unique value to medical applications. Motivated by the recent success of applying deep learning methods to medical image processing, we first propose an algorithmic architecture for supervised multimodal image analysis with cross-modality fusion at the feature learning level, classifier level, and decision-making level. We then design and implement an image segmentation system based on deep convolutional neural networks to contour the lesions of soft tissue sarcomas using multimodal images, including those from magnetic resonance imaging, computed tomography, and positron emission tomography. The network trained with multimodal images shows superior performance compared to networks trained with single-modal images. For the task of tumor segmentation, performing image fusion within the network (i.e., fusing at convolutional or fully connected layers) is generally better than fusing images at the network output (i.e., voting). This paper provides empirical guidance for the design and application of multimodal image analysis.

Index Terms—Computed tomography (CT), convolutional neural network (CNN), magnetic resonance imaging (MRI), multimodal image, positron emission tomography (PET).


providing quantitative metabolic and functional information about diseases can work together with CT and magnetic resonance imaging (MRI) which provide details on anatomic structures via high contrast and spatial resolution to better characterize lesions [2]. Another widely used multimodal imaging technique in neuroscience studies is the simultaneous recording of functional MRI (fMRI) and electroencephalography (EEG) [3], which offers both high spatial resolution (through fMRI) and temporal resolution (through EEG) on brain dynamics.




Correspondingly, various analyses using multimodal biomedical imaging and computer-aided detection systems have been developed. The premise is that various imaging modalities encompass abundant information which is different and complementary to each other. For example, in one deep-learning-based framework [4], automated detection of solitary pulmonary nodules were implemented by first identifying suspect regions from CT images, followed by merging them with high-uptake regions detected on PET images. As described in a multimodal imaging project for brain tumor segmentation [5], each modality reveals a unique type of biological/biochemical

❖ Evolution of “Alignment”

MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation

Cheng Chen ^a, Juzheng Miao ^b, Dufan Wu ^a, Aoxiao Zhong ^{d a}, Zhiling Yan ^c, Sekeun Kim ^a, Jiang Hu ^a, Zhengliang Liu ^{e a}, Lichao Sun ^c, Xiang Li ^a  , Tianming Liu ^e, Pheng-Ann Heng ^b, Quanzheng Li ^a

Show more 

 Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.media.2024.103310> 

[Get rights and content](#) 

Highlights

- A parameter-efficient adaptation of SAM to various volumetric and video medical data.
- A state-of-the-art segmentation performance on various medical imaging modalities.
- A comprehensive evaluation shows outstanding generalization capability of our model.
- A impressive performance improvement on tumor segmentation by leveraging prompts.

❖ Evolution of “Alignment”

Biomedical Visual Instruction Tuning with Clinician Preference Alignment

Hejie Cui^{1,2*}, Lingjun Mao^{3*}, Xin Liang³, Jieyu Zhang⁴,
Hui Ren^{5,6}, Quanzheng Li^{5,6}, Xiang Li^{5,6}, Carl Yang²

¹ Stanford University ² Emory University ³ University of California, Berkeley
⁴ University of Washington ⁵ Massachusetts General Hospital ⁶ Harvard Medical School

Abstract

Recent advancements in multimodal foundation models have showcased impressive capabilities in understanding and reasoning with visual and textual information. Adapting these foundation models trained for general usage to specialized domains like biomedicine requires large-scale domain-specific instruction datasets. While existing works have explored curating such datasets automatically, the resultant datasets are not explicitly aligned with domain expertise. In this work, we propose a data-centric framework, **Biomedical Visual Instruction Tuning with Clinician Preference Alignment** (BioMed-VITAL), that incorporates clinician preferences into both stages of generating and selecting instruction data for tuning biomedical multimodal foundation models. First, during the generation stage, we prompt the GPT-4V generator with a diverse set of clinician-selected demonstrations for preference-aligned data candidate generation. Then, during the selection phase, we train a separate selection model, which explicitly distills clinician and policy-guided model preferences into a rating function to select high-quality data for medical instruction tuning. Results show that the model tuned with the instruction-following data from our method demonstrates a significant improvement in open visual chat (18.5% relatively) and medical VQA (win rate up to 81.73%). Our instruction-following data and models are available at <https://BioMed-VITAL.github.io>.

❖ Evolution of “Alignment”

Eye-gaze Guided Multi-modal Alignment for Medical Representation Learning

Chong Ma, Hanqi Jiang, Wenting Chen, Yiwei Li, Zihao Wu, Xiaowei Yu, Zhengliang Liu, Lei Guo, Dajiang Zhu, Tuo Zhang, Dinggang Shen *Fellow, IEEE*, Tianming Liu *Senior Member, IEEE*, Xiang Li

Abstract—In the medical multi-modal frameworks, the alignment of cross-modality features presents a significant challenge. However, existing works have learned features that are implicitly aligned from the data, without considering the explicit relationships in the medical context. This data-reliance may lead to low generalization of the learned alignment relationships. In this work, we propose the Eye-gaze Guided Multi-modal Alignment (EGMA) framework to harness eye-gaze data for better alignment of medical visual and textual features. We explore the natural auxiliary role of radiologists’ eye-gaze data in aligning medical images and text, and introduce a novel approach by using eye-gaze data, collected synchronously by radiologists during diagnostic evaluations. We conduct downstream tasks of image classification and image-text retrieval on four medical datasets, where EGMA achieved state-of-the-art performance and stronger generalization across different datasets. Additionally, we explore the impact of varying amounts of eye-gaze data on model performance, highlighting the feasibility and utility of integrating this auxiliary data into multi-modal alignment framework.

Index Terms—Medical Multi-modal Alignment, Eye-gaze, Radiology.

between image and text data. For instance, GLIP [2] and RegionCLIP [3] utilized pre-predicted annotation information to perform fine-grained region-level pre-training. They introduced detection networks firstly to predict image regions relevant to the text prompt, and then trained the model to align these image regions with their corresponding text descriptions. However, these models heavily rely on the performance of the ROI detector and have high computational complexity. FILIP [4] proposed a refined multi-modal alignment operation after the encoder, relying solely on image patches and text tokens. Although this further explores the local feature relationships between multi-modal data, it still requires sufficient data support. When training on small-scale datasets, especially in the medical field, accurately learning alignment features between modalities becomes more challenging [5], [6].

To address the scarcity of medical data, studies [7], [8] have introduced self-supervised training into the CLIP framework to further enhance encoder performance. Additionally,

❖ Evolution of “Alignment”

REASONING BEFORE COMPARISON: LLM-ENHANCED SEMANTIC SIMILARITY METRICS FOR DOMAIN SPECIALIZED TEXT ANALYSIS

Shaochen Xu¹, Zihao Wu¹, Huaqin Zhao¹, Peng Shu¹, Zhengliang Liu¹, Wenxiong Liao², Sheng Li³, Andrea Sikora⁴, Tianming Liu¹, and Xiang Li⁵

¹School of Computing, University of Georgia

²School of Computer Science and Engineering, South China University of Technology

³School of Data Science, University of Virginia

⁴Department of Clinical and Administrative Pharmacy, University of Georgia College of Pharmacy

⁵Massachusetts General Hospital and Harvard Medical School

ABSTRACT

In this study, we leverage LLM to enhance the semantic analysis and develop similarity metrics for texts, addressing the limitations of traditional unsupervised NLP metrics like ROUGE and BLEU. We develop a framework where LLMs such as GPT-4 are employed for zero-shot text identification and label generation for radiology reports, where the labels are then used as measurements for text similarity. By testing the proposed framework on the MIMIC data, we find that GPT-4 generated labels can significantly improve the semantic similarity assessment, with scores more closely aligned with clinical ground truth than traditional NLP metrics. Our work demonstrates the possibility of conducting semantic analysis of the text data using semi-quantitative reasoning results by the LLMs for highly specialized domains. While the framework is implemented for radiology report similarity analysis, its concept can be extended to other specialized domains as well.

❖ Evolution of “Alignment”

Multimodal ChatGPT for Medical Applications: an Experimental Study of GPT-4V

Zhiling Yan^{1*} Kai Zhang^{1*} Rong Zhou¹ Lifang He¹ Xiang Li² Lichao Sun^{1†}

¹Lehigh University, ²Massachusetts General Hospital and Harvard Medical School

Abstract

In this paper, we critically evaluate the capabilities of the state-of-the-art multimodal large language model, i.e., GPT-4 with Vision (GPT-4V), on Visual Question Answering (VQA) task. Our experiments thoroughly assess GPT-4V’s proficiency in answering questions paired with images using both pathology and radiology datasets from 11 modalities (e.g. Microscopy, Dermoscopy, X-ray, CT, etc.) and fifteen objects of interests (brain, liver, lung, etc.). Our datasets encompass a comprehensive range of medical inquiries, including sixteen distinct question types. Throughout our evaluations, we devised textual prompts for GPT-4V, directing it to synergize visual and textual information. The experiments with accuracy score conclude that the current version of GPT-4V is not recommended for real-world diagnostics due to its unreliable and suboptimal accuracy in responding to diagnostic medical questions. In addition, we delineate seven unique facets of GPT-4V’s behavior in medical VQA, highlighting its constraints within this complex arena. The complete details of our evaluation cases are accessible at Github.

- ❖ Evolution of “Alignment”
- ❖ Generative modeling: advantages and pitfalls
- ❖ BiomedGPT: A “foundation” approach

Generative modeling: advantages and pitfalls

Fine-Grained Image-Text Alignment in Medical Imaging Enables Explainable Cyclic Image-Report Generation

Wenting Chen¹ Linlin Shen³ Jingyang Lin⁴ Jiebo Luo⁴

Xiang Li^{5*} Yixuan Yuan^{2*}

¹City University of Hong Kong ²The Chinese University of Hong Kong

³Shenzhen University ⁴University of Rochester

⁵Massachusetts General Hospital and Harvard Medical School

¹wentichen7-c@my.cityu.edu.hk ²xyxuan@ee.cuhk.edu.hk ³llshen@szu.edu.cn

⁴{jluo@cs, jlin81@ur}.rochester.edu ⁵xli60@mg.harvard.edu

Abstract

Fine-grained vision-language models (VLM) have been widely used for inter-modality local alignment between the predefined fixed patches and textual words. However, in medical analysis, lesions exhibit varying sizes and positions, and using fixed patches may cause incomplete representations of lesions. Moreover, these methods provide explainability by using heatmaps to show the general image areas potentially associated with texts rather than specific regions, making their explanations not explicit and specific enough. To address these issues, we propose a novel Adaptive patch-word Matching (AdaMatch) model to correlate chest X-ray (CXR) image regions with words in medical reports and apply it to CXR-report generation to provide explainability for the generation process. AdaMatch exploits the fine-grained relation between adaptive patches and words to provide explanations of specific image regions with corresponding words. To capture the abnormal regions of varying sizes and positions, we introduce an Adaptive Patch extraction (AdaPatch) module to acquire adaptive patches for these regions adaptively. Aiming to provide explicit explainability for the CXR-report generation task, we propose an AdaMatch-based bidirectional LLM for Cyclic CXR-report generation (AdaMatch-Cyclic). It employs AdaMatch to obtain the keywords for CXR images and 'keypatches' for medical reports as hints to guide CXR-report generation. Extensive experiments on two publicly available CXR datasets validate the effectiveness of our method and its superior performance over existing methods.

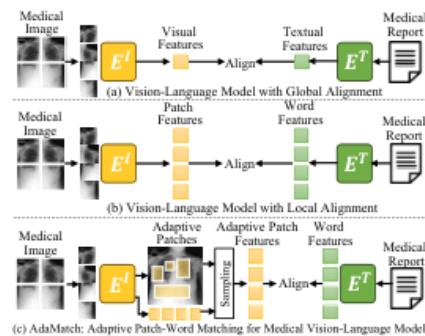


Figure 1: Current vision-language models (VLM) achieve (a) global alignment and (b) local alignment by matching overall visual with textual features, and aligning patches with word features, respectively. (c) To exploit the relation between textual words and abnormal patches with varied sizes, our AdaMatch obtains adaptive patch features and aligns them with word features.

learning (Radford et al., 2021). Technologies like contrastive learning and self-supervised learning have dramatically improved state-of-the-art alignment performance. Recent vision-language models (VLMs) demonstrate two approaches: global contrastive alignment, which integrates images and texts at a global level (Radford et al., 2021; Jia et al., 2021; Jang et al., 2023; Wang et al., 2023; Yang et al., 2022), and local alignment, focusing on detailed connections between visual objects and textual words (Chen et al., 2020a; Li et al., 2020b,a; Zhan et al., 2021; Kim et al., 2021; Yao et al., 2021),

❖ Generative modeling: advantages and pitfalls

High-Resolution 3d Ct Synthesis From Bidirectional X-Ray Images Using 3d Diffusion Model

Publisher: **IEEE**

[Cite This](#)

[PDF](#)

Siyeop Yoon ; Jay Sairam Pratap ; Wen-Chih Liu ; Matthew Tivnan ; Hui Ren ; Abhiram Bhashyam ; Quanzheng Li ; Neal Chen ; Xiang Li [All Authors](#)

16

Full

Text Views



Abstract

Document Sections

1. INTRODUCTION
2. SCORE BASED DIFFUSION MODELS
3. PATCH-WISE TRAINING OF DIFFUSION MODEL
4. EXPERIMENTS
5. RESULTS

Abstract:

3D Computed Tomography (CT) offers invaluable geometric insights into bone structures, but the high radiation dose and medical cost constraints are significant barriers. Moreover, CT reconstruction demands multiple X-ray projections, necessitating a dedicated scanning system, whereas bidirectional X-rays are already the front-line diagnostic tool in routine practice. Therefore, reconstructing 3D bone structures from bidirectional X-ray data can reduce the need for additional CT scans, provide rapid access to 3D information, and lower medical costs. Recently, diffusion models have emerged as potent tools for generating high-fidelity images. However, high computational costs have limited their utility. Collecting large-scale datasets for training in clinical environments presents another challenge. In this study, we introduce a novel approach to synthesize 3D CT volumes from a bidirectional X-ray projection using a 3D diffusion model. To reduce the computational burden and the need for a large dataset, our 3D diffusion model was trained using patch-wise loss. A conditional score function of our model incorporates 2D bidirectional X-ray images and patch coordinate information to synthesize high-resolution CT. Initial findings indicate that our diffusion model synthesizes 3D CT volumes from a bidirectional X-ray, effectively capturing 3D geometric correlations while enabling single-GPU training and rapid 3D volumetric sampling.

❖ Generative modeling: advantages and pitfalls

Zero-Shot Novel View Synthesis of Wrist X-Rays Using Latent Diffusion Model

Publisher: **IEEE**

[Cite This](#)

[PDF](#)

Jayanth Pratap ; Siyeop Yoon ; Wen-Chih Liu ; Quanzheng Li ; Abhiram Bhashyam ; Neal Chen ; Xiang Li [All Authors](#)

11

Full

Text Views



Abstract

Document Sections

1. INTRODUCTION
2. FORMULATION
3. METHODS
4. EXPERIMENTS AND RESULTS

Abstract:

X-ray imaging plays a crucial role in diagnosing and monitoring injuries such as distal radius fractures, which are among the most common musculoskeletal injuries. However, challenges such as patient comfort, radiation exposure, and cost/time constraints make it difficult to obtain a large number of X-ray views. We present a view-conditioned latent diffusion model capable of synthesizing new X-ray views of the wrist from a single X-ray input, enhancing the diagnostic capabilities and clinical utility of X-ray imaging. Preliminary results demonstrate the model's capability to generate realistic and clinically relevant X-ray views of the wrist from a single input, showing strong zero-shot performance on new patient anatomy and true radiographs.

Published in: [2024 IEEE International Symposium on Biomedical Imaging \(ISBI\)](#)

❖ Generative modeling: advantages and pitfalls

[Home](#) > [Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops](#) >

Conference paper

Graph-Based Counterfactual Causal Inference Modeling for Neuroimaging Analysis


Conference paper | First Online: 03 February 2024

pp 205–213 | [Cite this conference paper](#)



[Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops](#)

(MICCAI 2023)

[Haixing Dai](#), [Mengxuan Hu](#), [Qing Li](#), [Lu Zhang](#), [Lin Zhao](#), [Dajiang Zhu](#), [Ibai Diez](#), [Jorge Sepulcre](#), [Fan Zhang](#), [Xingyu Gao](#), [Manhua Liu](#), [Quanzheng Li](#), [Sheng Li](#), [Tianming Liu](#) & [Xiang Li](#) 

[Access this chapter](#)

- ❖ Evolution of “Alignment”
- ❖ Generative modeling: advantages and pitfalls
- ❖ BiomedGPT: A “foundation” approach