
MedTrinity-25M: A Large-scale Multimodal Dataset with Multigranular Annotations for Medicine

Yunfei Xie^{1,*} Ce Zhou^{1,*} Lang Gao^{1,*} Juncheng Wu^{2,*} Xianhang Li²
Hong-Yu Zhou³ Sheng Liu⁴ Lei Xing⁴ James Zou⁴ Cihang Xie² Yuyin Zhou²

*equal technical contribution

¹Huazhong University of Science and Technology ²UC Santa Cruz

³Harvard University ⁴Stanford University

Abstract

This paper introduces MedTrinity-25M, a comprehensive, large-scale multimodal dataset for medicine, covering over 25 million images across 10 modalities, with multigranular annotations for more than 65 diseases. These enriched annotations encompass both global textual information, such as disease/lesion type, modality, region-specific descriptions, and inter-regional relationships, as well as detailed local annotations for regions of interest (ROIs), including bounding boxes, segmentation masks. Unlike existing approach which is limited by the availability of image-text pairs, we have developed the first automated pipeline that scales up multimodal data by generating multigranular visual and textual annotations (in the form of image-ROI-description triplets) without the need for any paired text descriptions. Specifically, data from over 90 different sources have been collected, preprocessed, and grounded using domain-specific expert models to identify ROIs related to abnormal regions. We then build a comprehensive knowledge base and prompt multimodal large language models to perform retrieval-augmented generation with the identified ROIs as guidance, resulting in multigranular textual descriptions. Compared to existing datasets, MedTrinity-25M provides the most enriched annotations, supporting a comprehensive range of multimodal tasks such as captioning and report generation, as well as vision-centric tasks like classification and segmentation. Pretraining on MedTrinity-25M, our model achieves state-of-the-art performance on VQA-RAD and PathVQA, surpassing both multimodal large language models and other representative SoTA approaches. This dataset can also be utilized to support large-scale pre-training of multimodal medical AI models, contributing to the development of future foundation models in the medical domain. The dataset is publicly available at <https://yunfeixie233.github.io/MedTrinity-25M/>.

1 Introduction

Large-scale multimodal foundation models [1, 2, 3, 4, 5] have demonstrated remarkable success across various domains due to their ability to understand complex visual patterns in conjunction with natural language. This success has sparked significant interest in applying such models to medical vision-language tasks. Much progress has been made to improve the medical capacity of general domain multimodal foundation models by constructing medical datasets with image-text pairs and fine-tuning general domain models on these datasets [6, 7, 8, 9, 10].

However, current medical datasets have several limitations. Firstly, these datasets lack **multigranular** annotations that reveal the correlation between local and global information within medical images. Medical images often contain detailed cues, such as regional abnormal textures or structures, which

may indicate specific types of lesions. Therefore, multimodal models need the ability to infer global information, such as disease or lesion type, from local details. The absence of such data limits the models’ capacity to comprehensively understand medical images. Moreover, current dataset construction methods heavily rely on medical images paired with reports or captions, which restricts their scalability.

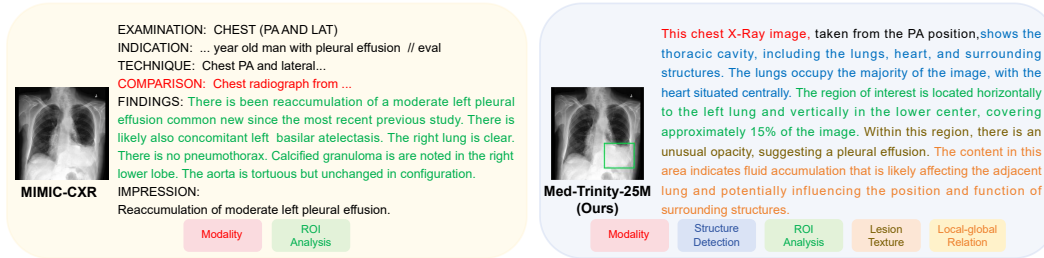
In this paper, we address the above challenges by proposing an automated data construction pipeline using multimodal large language models (MLLMs) without relying on paired text descriptions. To address the lack of comprehensive medical knowledge in general-purpose MLLMs, we leverage domain-specific expert grounding models and retrieval-augmented generation (RAG) to extract relevant medical knowledge. We then prompt MLLMs to generate multigranular visual and textual annotations enriched with this knowledge based on identified regions of interest (ROIs). We utilize this pipeline to transform the collected data, including large-scale unpaired images, into image-ROI-description triplets. These triplets provide multigranular annotations that encompass both global textual information, such as disease/lesion type, modality, and inter-regional relationships, as well as detailed local annotations for ROIs, including bounding boxes, segmentation masks, and region-specific textual descriptions. Using the proposed pipeline, we create a large-scale multimodal multigranular medical dataset containing over 25 million triplets, named **MedTrinity-25M**. To our best knowledge, this is the largest multimodal dataset in medicine to date.

Initially, we assemble a large amount of medical data from over 90 online resources such as TCIA, Kaggle, Zenodo, Synapse, etc. In addition to images with a small amount of high-quality paired manual reports, this assembled data also includes two types of coarse medical data: 1) Image data with segmentation masks, lesion bounding boxes, or only disease types but lacking detailed textual descriptions, and 2) Images paired with coarse captions that describe only global modality or disease information, but lack detailed descriptions of local regions. To generate multigranular annotations from the massive coarse medical data, we first identify ROIs that contain disease or lesion patterns by applying expert grounding models. We then build a comprehensive knowledge base from online corpora (e.g., PubMed) and retrieve image-related medical knowledge. Finally, we prompt MLLMs to integrate medical knowledge with guidance of identified ROIs to generate multigranular textual descriptions.

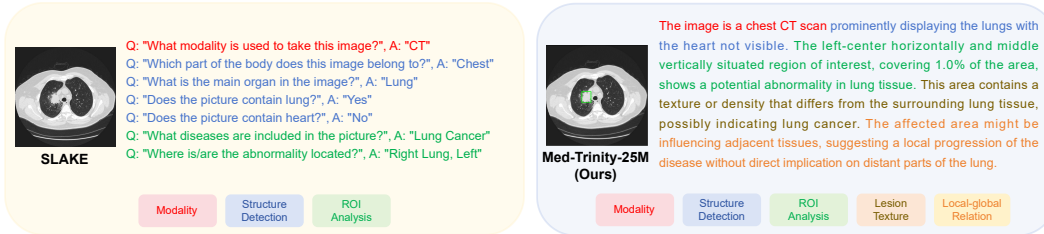
2 Related Work

Medical Multimodal Foundation Models. Due to the effectiveness of multimodal foundation models in understanding visual features, adapting these models to perform medical vision-language tasks has garnered increasing attention in recent years [11, 12, 9, 5]. Several papers attempt to adapt general domain multimodal foundation models with varying architecture to medical domain through end-to-end training on medical datasets. For example, Med-Flamingo [11] enhances the medical capacity of OpenFlamingo-9B [13] by fine-tuning it with 0.8M interleaved and 1.6M paired medical image-text data. While Med-PaLM [12] adapts PaLM-E [14] to medical domain using approximately 1M medical data points, demonstrating competitive or surpassing performance compared to state-of-the-art models. Additionally, LLaVA-Med [9] employs end-to-end visual instruction tuning [1] with two stages, achieving remarkable results in medical Visual Question Answering (VQA) tasks. Similarly, Med-Gemini [15] employs a long-form question answering dataset to enhance the multimodal and long-context capabilities of baseline Gemini [16]. Although these models have achieved remarkable performance, they are still limited by the scale of training data. Prior research [17] has shown that scaling up the training data improves the performance of large multimodal foundation models. In this paper, we aim to build a large-scale medical dataset to facilitate the development of more powerful medical multimodal foundation models.

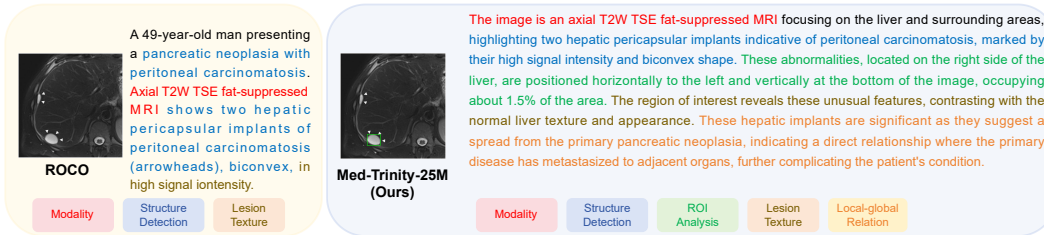
Multimodal Datasets for medicine. The significance of constructing comprehensive medical multimodal datasets has garnered considerable attention [9, 18, 19, 7]. Several works attempt to collect images and paired clinical reports prepared by pathology specialist [19, 7, 8], which provide comprehensive descriptions of images, including disease types and corresponding reasoning. For example, MIMIC-CXR[8] comprises 227,835 images for 65,379 patients, containing pathological findings and impressions in reports paired with each images. However, manually constructing such reports is both time-consuming and expensive, thereby limiting the scale of these datasets. PMC-OA [20] aims to expand the dataset scale by extracting a large number of image-caption pairs from medical papers, increasing the number of data samples to 1.65 million. However, the extracted



(a) Qualitative Comparison with sample in radiology report of chest x-rays dataset MIMIC-CXR [21].



(b) Qualitative Comparison with sample in visual QA dataset SLAKE [22].



(c) Qualitative Comparison with sample in radiology objects caption dataset ROCO [18].

Figure 1: Qualitative comparison with different types of dataset.

captions are less detailed compared to manual clinical reports, resulting in a lack of multigranular annotations. RadGenome-Chest CT [19] includes more detailed annotations, such as segmentation masks and medical reports generated by MLLMs. Nonetheless, its construction method still relies on paired image-text data, which limits its scalability. Unlike these existing methods, we devise the first automated data construction pipeline to generate multigranular annotations for unpaired images, achieving a comprehensive multigranular dataset with 25 million data samples.

3 MedTrinity-25M Dataset

3.1 Data Triplet

Our dataset comprises triplets of $\{\text{image}, \text{ROI}, \text{description}\}$. Each ROI is associated with an abnormality and is represented by a bounding box or a segmentation mask, specifying the relevant region within the image. For each image, we provide a multigranular textual description, which includes the disease/lesion type, modality, region-specific description, and inter-regional relationships as illustrated in Figure 2.

Images. We use the original medical image in the source dataset, we extensively collected medical datasets from the following sources: (1) online resources such as TCIA, Kaggle, Zenodo, Synapse, Hugging Face, Grand Challenge, GitHub, etc. (2) relevant medical dataset research, such as CheXpert [7] and DeepLesion [23]. These datasets were first categorized into two types: (1) datasets containing local annotations, such as MIMIC-CXR [8] with corresponding radiology reports, and PMC-OA [24] with corresponding captions, where the reports or captions provide analysis of specific local conditions in the images; another example is the 3D image segmentation dataset BraTS2024 [25], which marks the tumor regions in CT scans with masks. (2) datasets containing global annotations: such as image classification datasets ISIC2019 [26] and ISIC2020 [27], whose classification labels

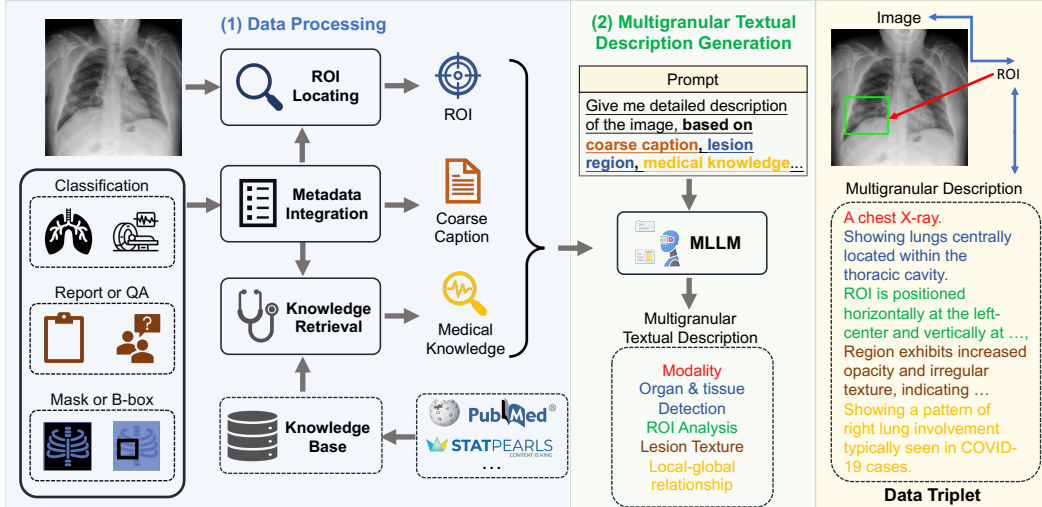


Figure 2: **Data construction pipeline.** 1) Data processing: extracting essential information from collected data, including **metadata integration** to generate coarse caption, **ROI locating**, and **medical knowledge collection**. 2) Multigranular textual description generation: using this information to prompt MLLMs to generate fine-grained captions.

reflect the overall pathological condition of tissue sections; another example is the CheXpert [7] dataset, which provides detailed classification of disease types for each chest X-ray. We collect 25,001,668 samples spanning 10 modalities and over 65 diseases. For 3D volumetric images stored in DICOM or NIfTI formats, we converted each 2D slice to PNG format. Additional caption and annotations like masks and bounding boxes from these datasets were utilized to construct ROIs and corresponding textual descriptions as below.

ROIs. For each image, ROIs are highlighted using segmentation masks or bounding boxes. These ROIs mostly contain pathological findings such as lesions, inflammation, neoplasms, infections, or other potential abnormalities. In the few cases without abnormalities, the ROIs generally indicate the primary object or organ in the image, as shown in examples in the supplementary material.

Textual Descriptions. The textual descriptions for each image are provided with detailed information across various aspects. Unlike the unstructured free-text descriptions found in previous medical report datasets[7, 8, 6] or simple short sentences in visual QA dataset[28, 22] and caption dataset[18, 24], our textual descriptions are multigranular and structured. General attributes related to the image are described first, including the image modality, the specific organ depicted, and the type of disease presented. Subsequently, ROI-related information is provided, including their locations and the abnormal characteristics within them that indicate underlying pathology, such as distinctive color and texture. Additionally, comparisons between the ROIs and surrounding regions are presented to highlight differences in features and the extent of disease progression.

We also demonstrate the multigranular textual descriptions in our dataset with those in other common forms. As illustrated in Figure 1, our textual description is multigranular with more attributes than radiology report of chest x-rays dataset MIMIC-CXR [21], visual QA dataset SLAKE[22] and radiology objects caption dataset ROCO[18].

3.2 Data Construction Pipeline

Given a medical image, we aim to generate corresponding multigranular visual and textual annotations by leveraging MLLMs. Specifically, as shown in Figure 2, our pipeline can be decomposed into two stages - **Data Processing** and **Generation of Multigranular Text Description**. In the **Data Processing** stage (Section 3.2.1), we address the lack of domain-specific knowledge in general-purpose MLLMs by leveraging expert grounding models and retrieval-augmented generation (RAG). This stage includes three key steps: 1) **Metadata Integration** to produce coarse captions encapsulating fundamental image information such as modality and disease types; 2) **ROI Locating** to identify

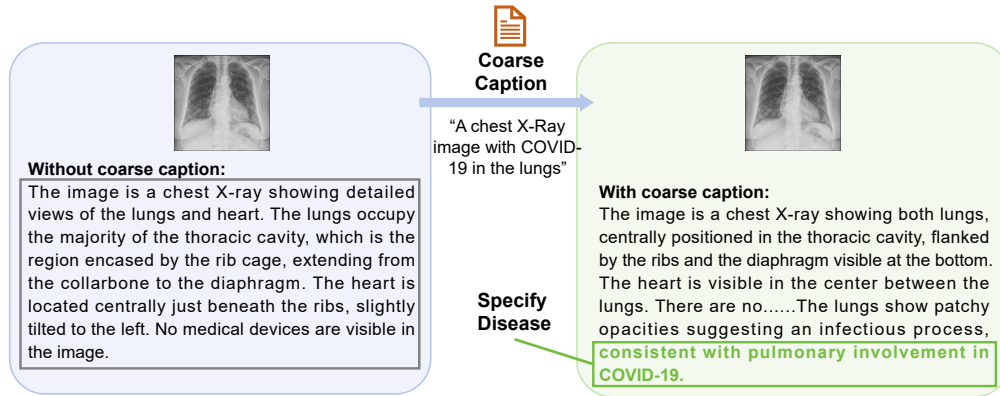


Figure 3: A qualitative comparison example of generated textual description with and without coarse caption. Without a coarse caption, MLLMs fails to detect diseases. On the contrary, providing a caption mentioning “COVID-19” allows MLLMs to identify and categorize the disease, facilitating further analysis.

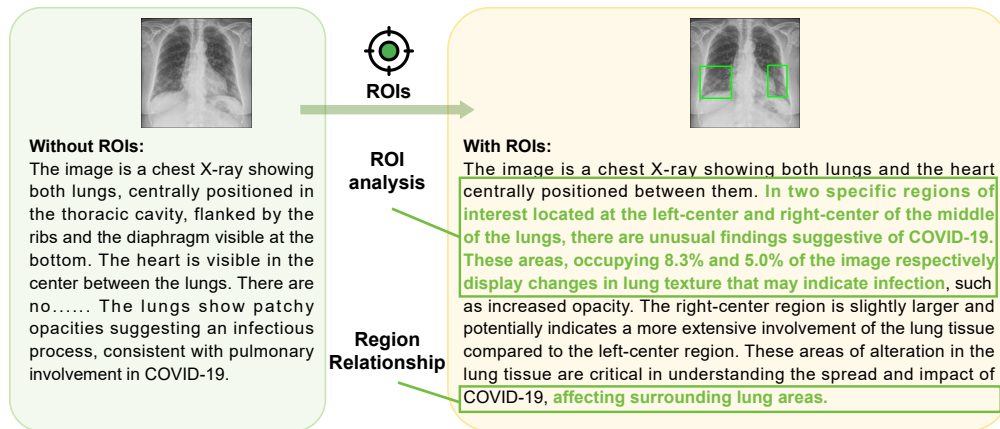


Figure 4: A qualitative comparison example of generated textual description with and without locating ROIs. Without ROIs, the caption offers only a brief global analysis; with ROIs, MLLMs conducts detailed local analysis and assesses the impact of lesion ROIs on adjacent normal regions.

regions of abnormalities; and 3) **Medical Knowledge Retrieval** to extract relevant fine-grained medical details. Based on the processed data, we then prompt MLLMs to generate multigranular text descriptions, resulting in the creation of fine-grained captions, as detailed in Section 3.2.2.

3.2.1 Data Processing

Coarse Caption Generation via Metadata Integration. We aim to generate coarse captions that provide fundamental information for a given image, including modality, organ labels, disease types, and optionally, camera views and equipment information. Instead of extracting features directly from the images, we generate these captions by integrating dataset metadata. We first extract metadata from the datasets and then apply a fixed rule to integrate this information into coarse captions. For example, for an image from the QaTa-COV19 dataset¹, we derive metadata from the dataset’s accompanying paper or documentation, indicating that it consists of COVID-19 chest X-ray images. Next, we construct coarse captions like “A chest X-ray image with COVID-19 in the lungs” highlighting the modality, organ types, and disease labels. If the image contains additional textual information like radiological findings, this is also integrated to enhance the richness of the caption. The effectiveness of adding coarse captions when generating fine-grained captions is illustrated in Figure 3. In contrast to the scenario without a coarse caption where MLLMs fails to recognize the disease, providing

¹<https://www.kaggle.com/aysendegerli/qatacov19-dataset>.

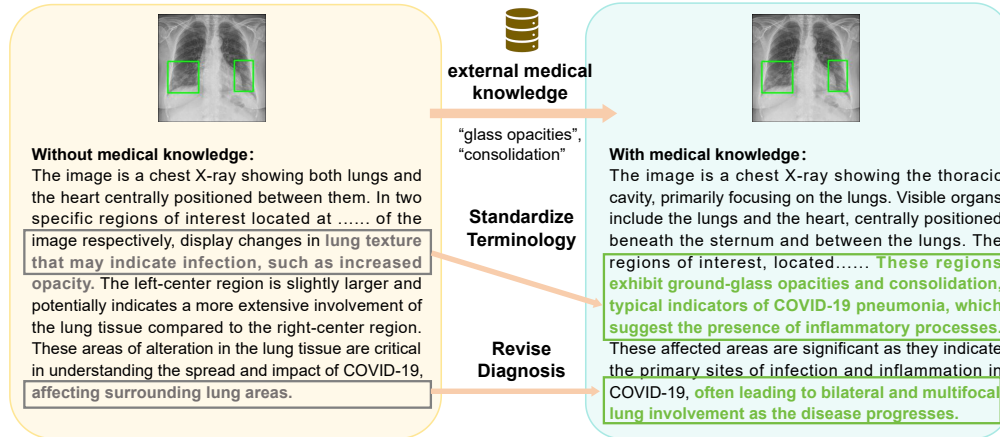


Figure 5: A qualitative comparison example of generated textual description with and without external medical knowledge. MLLMs can standardize medical terminology in its expressions and refine its diagnosis based on disease progressions detailed in medical literature.

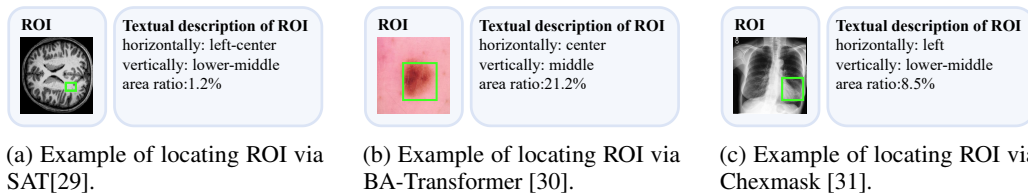


Figure 6: Example of ROIs and their corresponding textual descriptions.

MLLMs with a coarse caption that includes the disease type “COVID-19” enables it to identify and categorize the disease, thereby laying the foundation for further analysis.

ROI Locating. We employ various strategies to locate Regions of Interest (ROIs) in images. For datasets that already include localization annotations, such as segmentation masks or bounding boxes, we derive the ROIs from these existing annotations. Specifically, bounding boxes are directly used as the ROIs, while segmentation masks are converted to ROIs by creating the smallest bounding box that covers the mask. When such localization annotations are not available, we apply different pretrained expert models listed in the Appendix to generate ROIs. For text-prompt driven grounding model[29], we use disease and organ information in coarse captions as text prompts to guide the model in segmenting specific parts. Examples of generated ROIs from various modalities with different models are demonstrated in Figure 6. It is important to note that for modalities such as X-ray and MRI scans viewed from the z-axis, our ROI localization employs a coordinate system relative to the human body, resulting in a left-right reversal in the image representation.

Without ROIs, the original description is limited to a brief global analysis of the image. However, with ROIs, MLLMs can perform a more detailed local analysis of the ROIs and assess the impact of lesion ROIs on the surrounding normal regions, as demonstrated in Figure 4.

Medical Knowledge Retrieval. General-purpose MLLMs often produce content that lacks specialized medical terminology and professional expression. To address this issue, we build a medical knowledge database following the approach in MedRAG [32]. We collect three main corpora: PubMed² for biomedical knowledge, StatPearls³ for clinical decision support, and medical textbooks [33] for domain-specific knowledge. We segment these corpora into short snippets and encode them into high-dimensional vectors using the text encoder from Med-CPT [34]. These vectors are then indexed into a specialized vector knowledge base using Faiss[35], optimized for efficient retrieval.

²<https://pubmed.ncbi.nlm.nih.gov/>

³<https://www.statpearls.com/>

Knowledge 1:**Title: Mobile chest X-ray manifestations of 54 deceased patients with coronavirus disease 2019: Retrospective study.**

Content: We found that 50 (93%) patients with **lesions occurred in the bilateral lung**, 4 (7%) patients occurred in the right lung, 54 (100%) patients were **multifocal involvement**. The number of lung fields involved was 42 (78%) patients in 6 fields, 3 (6%) patients in 5 lung fields, 4 (7%) patients in 4 lung fields, and 5 (9%) patients in 3 lung fields. Fifty-three (98%) patients had **patchy opacities**, 3 (6%) patients had round or **oval solid nodules**, 9 (17%) patients had fibrous stripes, 13 (24%) patients had **pleural effusion**, 8 (15%) patients had **pleural thickening**, 6 (11%) patients had **pneumothorax**, 3 (6%) patients had **subcutaneous emphysema**. Among the 24 patients who had serial mobile chest X-rays, 16 (67%) patients had the progression of the lesions, 8 (33%) patients had no significant change of the lesions, and there was no case of reduction of the lesions. The mobile chest X-ray manifestations of deceased patients with COVID-19 were **mostly bilateral lung, multifocal involvement, and extensive lung field, and pleural effusion, pleural thickening, and pneumothorax probably could be observed**. The serial mobile chest X-ray showed that the chest lesions were progressive with a high probability.

.....

Figure 7: **An example of the Top-8 retrieval results.** By leveraging COVID-19-related medical knowledge, MLLMs can standardize medical terminology and enhance diagnoses according to the disease progressions described in medical literature.

For a given image, we retrieve relevant medical knowledge by using its coarse caption, which is generated through metadata integration. Specifically, we encode the coarse captions, including disease and organ classifications, into vectors using the Med-CPT text encoder. We then perform a vector similarity search in the medical vector database, retrieving the top eight medical knowledge snippets that semantically match the query. These snippets provide the external medical knowledge paired with the image. A qualitative example demonstrating the effectiveness of incorporating external medical knowledge is shown in Figure 7. With access to COVID-19-related medical knowledge, MLLMs can standardize medical terminology and refine diagnoses based on the disease progressions outlined in medical literature.

3.2.2 Generation of Multigranular Text Description

After data processing, a comprehensive prompt is utilized to guide the MLLMs in generating multigranular descriptions. The prompt template consists of a three-level hierarchical framework with questions to instruct MLLMs: (1) a global description that captures all details of the image (2) a local-focused analysis of specific ROIs that potentially are unusual; and (3) a local-global examination of the interaction between local and global attributes to understand the impact of local abnormalities on the entire organ. Detailed prompt template is presented in supplementary materials.

To ensure that the MLLMs are guided by relevant medical information not inherently present in their training data, we incorporate the processed data (coarse captions, ROIs, and retrieved medical knowledge) into the prompts. Specifically, for global information, coarse captions are directly integrated into the prompt. For local information, ROIs on images are converted into textual descriptions based on their coordinates and area ratio within the images. Examples of these textual descriptions are shown in Figure 6, using terms such as “left-center” and “area ratio: 1.2%”.

To refine terminology and diagnosis within ROIs, relevant medical knowledge about specific diseases is incorporated into the prompt. Instead of merely inserting this knowledge, we instruct MLLMs to identify and align the relevant knowledge to ROIs that require analysis.

Choice of MLLMs We first prompt GPT-4V with the provided medical coarse captions, ROIs, and medical knowledge to generate a subset of 200,000 samples, maintaining a similar modality and organ distribution to our full 25 million dataset. The goal of curating this subset is to calibrate a medical knowledge-guided MLLM to adhere to the formatting instructions specified for our text. Subsequently, we employ our model, LLaVA-Med Captioner, which is based on LLaVA-Med [9], the state-of-the-art medical MLLM. To further improve this model, we leverage the latest LLaMA3[36] to enhance its linguistic capabilities, and incorporate multi-scale feature extraction [37] to improve its vision capabilities. LLaVA-Med Captioner undergoes continuous training on medical multimodal data and is fine-tuned using our multigranular annotations, resulting in a specialized medical model.

After fine-tuning, we then use this specialized model to generate the multigranular text descriptions on our entire dataset, resulting in 25 million image-ROI-description triplets. The fine-tuning process leverages the advanced language organization capabilities of GPT-4V, providing an effective template for fine-grained captions, which our model uses to learn the formatting of fine-grained captions. As a result, our model generates more detailed descriptions compared to GPT-4V, as illustrated in Figure 8. We also show a detailed quantitative comparison in appendix B in the supplementary material.

3.3 Dataset Analysis

Diversity Our dataset encompasses a wide range of 10 imaging modalities, with more than 65 diseases across various anatomical structures in human. The distribution of Anatomical and biological

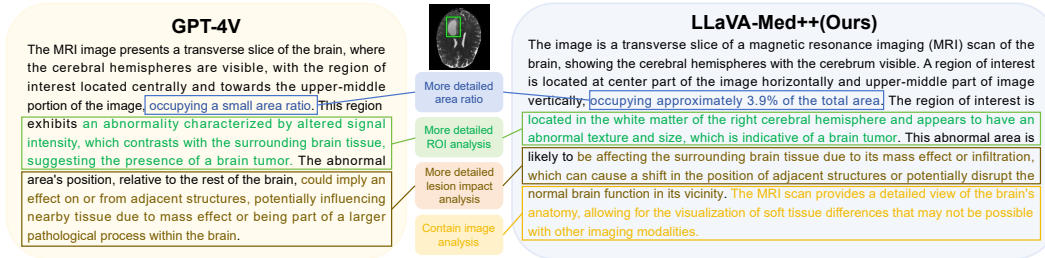
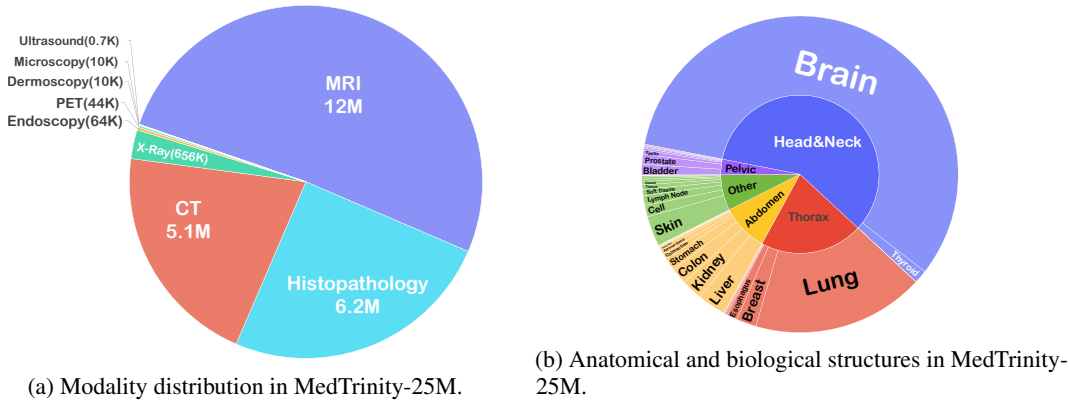
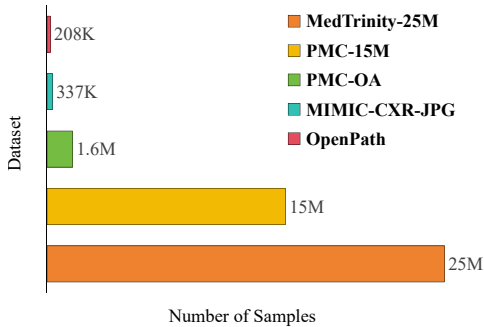
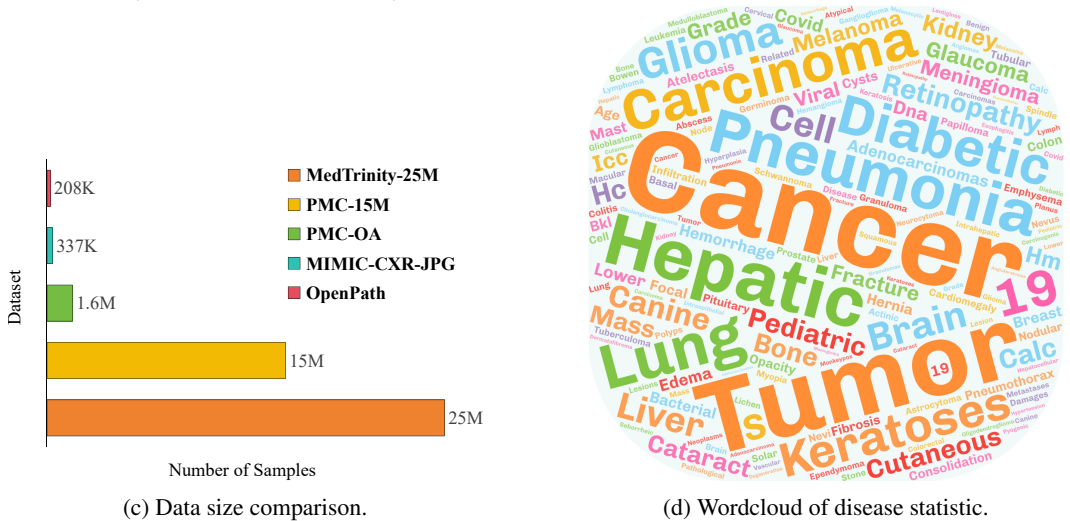


Figure 8: **Qualitative Comparison** with sample generated by GPT-4V. Compared to GPT-4V, our model generate more detailed caption.



(a) Modality distribution in MedTrinity-25M.

(b) Anatomical and biological structures in MedTrinity-25M.



(c) Data size comparison.

(d) Wordcloud of disease statistic.

Figure 9: Statistical overview of MedTrinity-25M.

structures in MedTrinity-25M is shown in Figure 9b. Meanwhile, the number of samples in the dataset for each modality are shown in Figure 9a, spanning from common ones with over 1 million samples each (CT, MRI, X-ray) to rare modalities(ultrasound, dermoscopy), demonstrating a much more balanced distribution compared to other large-scale dataset like SA-Med2D-20M[38], which only contain thousands of ultrasound and dermoscopy samples.

Scale Figure 9c shows the amount of our dataset, which is significantly larger than previous datasets. To the best of our knowledge, this is the largest open-source, multi-modal multigranular medical dataset to date.

Diseases The datasets involved in constructing MedTrinity-25M primarily focus on disease diagnosis and medical discovery. In MedTrinity-25M, diseases are given in the free-form text. The same disease may be referred to using different terms, allowing for elaborate identification and analysis. Figure 9d illustrates the frequently used words related to diseases in our dataset.

Dataset	Modality	Lesion Type	Lesion BBox/Mask	Color Texture Description	Region Relationship
MedMNIST [39]	✗	✓	✗	✗	✗
DeepLesion [40]	✓	✗	✓	✗	✗
BraTS 2024 [41]	✓	✗	✓	✗	✗
MIMIC-CXR [21]	✓	✓	✓	✓	✗
Quilt-1M [10]	✓	✓	✗	✓	✓
VQA-RAD [42]	✓	✓	✗	✓	✗
CRC100K [43]	✓	✓	✗	✗	✗
SA-Med2D-20M [44]	✓	✓	✓	✗	✗
MedTrinity-25M(Ours)	✓	✓	✓	✓	✓

Table 1: Comparison of dataset types based on provided attributes of annotations.

Table 2: Comparison of alignment scores between our generated fine-grained captions and human annotations.

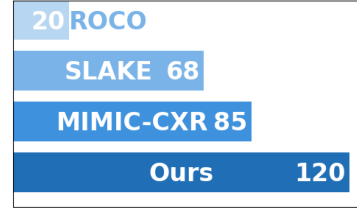


Figure 10: Comparison of the average word count of text descriptions.

(a) Alignment Scores on SLAKE

Score	SLAKE					
	Overall	Modality	Structure Detection	ROI Analysis	Lesion Texture	Local-Global Relation
Ours	8.2/10.0	2.0/2.0	1.7/2.0	1.8/2.0	1.6/2.0	1.1/2.0

(b) Alignment Scores on MIMIC-CXR

Score	MIMIC-CXR					
	Overall	Modality	Structure Detection	ROI Analysis	Lesion Texture	Local-Global Relation
Ours	8.9/10.0	2.0/2.0	1.9/2.0	1.8/2.0	1.6/2.0	1.6/2.0

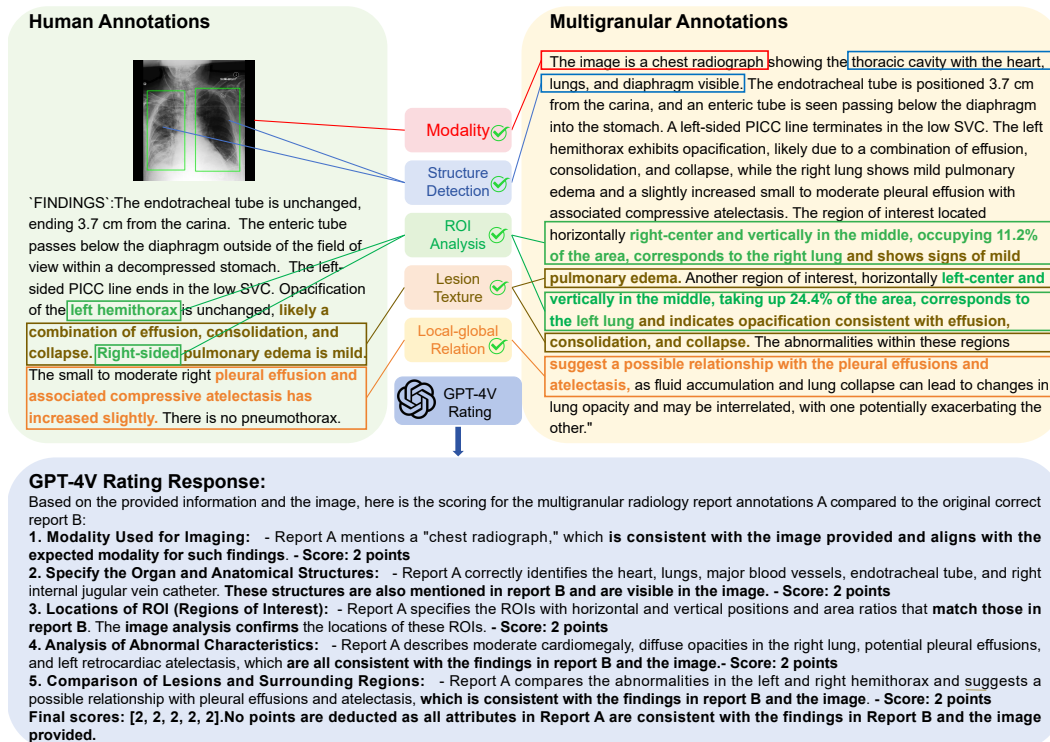
Richness We provide both quantitative analysis and qualitative examples to show the richness of our generated multigranular compare to other medical dataset. Qualitative examples are shown in Figure 1, our textual description is multigranular with more attributes than radiology report of chest x-rays dataset MIMIC-CXR [21], visual QA dataset SLAKE[22] and radiology objects caption dataset ROCO[18]. To demonstrate the multi-granularity of our data, we compared the average word count of text descriptions in our dataset, MedTrinity-25M, with those in other medical datasets, as illustrated in Figure 10. The word count in our dataset is significantly higher, indicating greater richness.

Alignment with human To evaluate the validity and quality of the generated multigranular annotations, we compared them with their original human annotations to assess the degree of alignment (for samples with human annotations).

Since the generated fine-grained captions contains structured descriptions that may significantly differ from free-text radiology reports and question-answering pairs, we leveraged GPT-4V’s vision and language understanding capabilities. Rather than focusing on the exact alignment of sentence structure or organization, GPT-4V assessed the alignment based on the accuracy of medical facts and diagnoses. Specifically, the structure of the generated fine-grained captions consists of five key attributes that characterize a medical image: modality, structure detection, ROI analysis, lesion texture, and local-global relation. To evaluate the generated data, we had GPT-4V perform a detailed comparison with human annotations based on these five attributes. Each attribute was scored on a scale from 0 to 2 points, with a maximum possible total score of 10 points.

We conducted an alignment study on SLAKE [22] and MIMIC-CXR [21], randomly selecting 50 samples to compare with fine-grained captions for evaluating alignment scores against human annotations. As shown in Table 2, the alignment scores were 8.2 and 8.9 for SLAKE and MIMIC-CXR, respectively. The criteria of modality, structure detection, and ROI analysis nearly achieved perfect scores, demonstrating the validity and accuracy of the generated data compared to human

Figure 11: An example of a perfect score result evaluated by GPT-4V. GPT-4V assesses five criteria, each fully aligned with human annotations, resulting in perfect scores.



annotations. An example of perfect alignment score results evaluated by GPT-4V is shown in Figure 11. In these examples, GPT-4V fully aligned with human annotations across all five criteria, resulting in perfect alignment scores.

The prompt used to query GPT-4V for evaluating the alignment score is shown in Figure 14 of supplementary.

4 LLaVA-Med++: Experimental Training with MedTrinity-25M

To further demonstrate the validity of our dataset, we compare the performance of LLaVA-Med++ with and without training on our dataset. We select Visual Question Answering (VQA) as the evaluation task, which requires models to learn detailed visual and language representations. We assessed the performance of our model on three biomedical VQA datasets: VQA-RAD [42], SLAKE [22] and PathVQA [45].

We initially pretrained LLaVA-Med++ using the methodology of LLaVA-Med [9] as our baseline. Subsequently, for each VQA dataset evaluation, we further pretrained our model on the corresponding MedTrinity-25M subset to achieve multigranular alignment. The model was then fine-tuned on VQA datasets for three epochs, with performance results presented in Table 3. A comparative experiment was conducted without pretraining on MedTrinity-25M, maintaining all other settings. Results clearly demonstrate that LLaVA-Med++ achieves state-of-the-art performance in two of the three VQA benchmarks and ranks third in the remaining one. Pretraining on MedTrinity-25M exhibits performance improvements of approximately 10.75% on VQA-RAD, 6.1% on SLAKE, and 13.25% on PathVQA compared to the model trained without pretraining on it. This enhancement underscores the efficacy of pretraining on MedTrinity-25M for downstream multimodal medical tasks, particularly in visual question answering.

Method	VQA-RAD			SLAKE			PathVQA		
	Ref	Open	Closed	Ref	Open	Closed	Ref	Open	Closed
<i>Supervised finet-tuning results with our own experiment runs</i>									
GPT-4V [2]		39.5	78.9		33.6	43.6		-	-
LLaVA		50.0	65.1		78.2	63.2		7.7	63.2
LLaVA-Med		55.5	66.5		70.6	54.5		35.9	89.2
LLaVA-Med++(Ours, w/o)		64.6	77.0		79.3	84.0		55.0	94.0
LLaVA-Med++(Ours, w/)		77.1 (+12.5)	86.0 (+9.0)		86.2 (+6.9)	89.3 (+5.3)		66.5 (+11.5)	99.0 (+5.0)
<i>Representative & SoTA methods with numbers reported in the literature</i>									
VL Encoder-Decoder [46]		71.5	82.5					71.5	85.6
Q2ATransformer [47]		79.2	81.2					54.9	88.9
Prefix T. Medical LM [48]				84.3		82.0		40.0	87.0
PubMedCLIP [49]		60.1	80.0		78.4	82.5			
BiomedCLIP [50]		67.6	79.8		82.1	89.7			
M2I2 [51]		66.5	83.5		74.7	91.1		36.3	88.0

Table 3: **Comparison with Existing Supervised Methods.**The notation **w/** and **w/o** indicate models with and without pretraining on MedTrinity-25M, respectively. Employing multigranular alignment pretraining on MedTrinity-25M, LLaVA-Med++ achieves state-of-the-art performance in two of the three VQA benchmarks and ranks third in the remaining one. Our model surpasses both multimodal large language models and other representative SoTA approaches.

5 Conclusion

This paper introduces MedTrinity-25M, a large-scale multimodal medical dataset comprising over 25 million image-ROI-description triplets sourced from more than 90 online resources, spanning 10 modalities and covering over 65 diseases. Unlike existing dataset construction methods that rely on image-text pairs, we have developed the first automated pipeline to scale up multimodal data by generating multigranular visual and textual annotations from unpaired image inputs, leveraging expert grounding models, retrieval-augmented generation techniques, and advanced MLLMs. MedTrinity-25M’s enriched annotations have the potential to support a wide range of multimodal tasks, such as captioning, report generation, classification, and segmentation, as well as facilitate the large-scale pre-training of multimodal medical AI models.

Acknowledgement

We thank the Microsoft Accelerate Foundation Models Research Program, the OpenAI Researcher Access Program, TPU Research Cloud (TRC) program, Google Cloud Research Credits program, AWS Cloud Credit for Research program, and Lambda Cloud for supporting our computing needs.

References

- [1] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [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.
- [3] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.
- [4] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [5] Hong-Yu Zhou, Subathra Adithan, Julián Nicolás Acosta, Eric J Topol, and Pranav Rajpurkar. A generalist learner for multifaceted medical image interpretation. *arXiv preprint arXiv:2405.07988*, 2024.
- [6] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

- [7] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.
- [8] Alistair EW 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. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.
- [9] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.
- [10] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in Neural Information Processing Systems*, 36, 2024.
- [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.
- [12] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.
- [13] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.
- [14] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, pages 8469–8488. PMLR, 2023.
- [15] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.
- [16] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [17] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [18] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer, 2018.
- [19] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*, 2024.
- [20] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.
- [21] AlistairEW Johnson, TomJ Pollard, SethJ Berkowitz, NathanielR Greenbaum, MatthewP Lungren, Chih-ying Deng, RogerG 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.

- [22] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.
- [23] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations. *arXiv preprint arXiv:1710.01766*, 2017.
- [24] axiong/pmc_oa datasets at hugging face. https://huggingface.co/datasets/axiong/pmc_oa.
- [25] Alexandros Karargyris, Renato Umeton, Micah J Sheller, Alejandro Aristizabal, Johnu George, Anna Wuest, Sarthak Pati, Hasan Kassem, Maximilian Zenk, Ujjwal Baid, et al. Federated benchmarking of medical artificial intelligence with medperf. *Nature Machine Intelligence*, 5(7):799–810, 2023.
- [26] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.
- [27] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):34, 2021.
- [28] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.
- [29] Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. One model to rule them all: Towards universal segmentation for medical images with text prompts. *arXiv preprint arXiv:2312.17183*, 2023.
- [30] Jiacheng Wang, Lan Wei, Liansheng Wang, Qichao Zhou, Lei Zhu, and Jing Qin. Boundary-aware transformers for skin lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 206–216. Springer, 2021.
- [31] Nicolas Gaggion, Lucas Mansilla, Candelaria Mosquera, Diego H. Milone, and Enzo Ferrante. Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis. *IEEE Transactions on Medical Imaging*, 2022.
- [32] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*, 2024.
- [33] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- [34] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.
- [35] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.
- [36] Meta LLaMA Team. Introducing meta llama 3: The most capable openly available llm to date. <https://ai.meta.com/blog/meta-llama-3/>, 2024.
- [37] Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When do we not need larger vision models? *arXiv preprint arXiv:2403.13043*, 2024.
- [38] Jin Ye, Junlong Cheng, Jianpin Chen, Zhongying Deng, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyang Huang, Jilong Chen, Lei Jiang, et al. Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. *arXiv preprint arXiv:2311.11969*, 2023.

- [39] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.
- [40] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations. *arXiv preprint arXiv:1710.01766*, 2017.
- [41] Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwal Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, et al. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. *arXiv preprint arXiv:2405.18368*, 2024.
- [42] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.
- [43] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. <https://doi.org/10.5281/zenodo.1214456>.
- [44] Jin Ye, Junlong Cheng, Jianpin Chen, Zhongying Deng, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyang Huang, Jilong Chen, Lei Jiang, et al. Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. *arXiv preprint arXiv:2311.11969*, 2023.
- [45] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.
- [46] Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. Vision-language model for visual question answering in medical imagery. *Bioengineering*, 10(3):380, 2023.
- [47] Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. Q2atransformer: Improving medical vqa via an answer querying decoder. In *International Conference on Information Processing in Medical Imaging*, pages 445–456. Springer, 2023.
- [48] Tom Van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees GM Snoek, and Marcel Worring. Open-ended medical visual question answering through prefix tuning of language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 726–736. Springer, 2023.
- [49] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, 2023.
- [50] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.
- [51] Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. Self-supervised vision-language pretraining for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.
- [52] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.
- [53] Eduardo Pontes Reis, Felipe Nascimento, Mateus Aranha, Fernando Mainetti Seicol, Birajara Machado, Marcelo Felix, Anouk Stein, and Edson Amaro. Brain hemorrhage extended (bhx): Bounding box extrapolation from thick to thin slice ct images. *PhysioNet*, 101(23):e215–20, 2020.
- [54] Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwal Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, et al. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. *arXiv preprint arXiv:2405.18368*, 2024.
- [55] Alexandros Karargyris, Renato Umeton, Micah J Sheller, Alejandro Aristizabal, Johnu George, Anna Wuest, Sarthak Pati, Hasan Kassem, Maximilian Zenk, Ujjwal Baid, et al. Federated

- benchmarking of medical artificial intelligence with medperf. *Nature Machine Intelligence*, 5(7):799–810, 2023.
- [56] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7(1):29, 2016.
- [57] Angel Cruz-Roa, Ajay Basavanahally, Fabio González, Hannah Gilmore, Michael Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, and Anant Madabhushi. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *Medical Imaging 2014: Digital Pathology*. SPIE, March 2014.
- [58] Kexin Ding, Mu Zhou, He Wang, Olivier Gevaert, Dimitris Metaxas, and Shaoting Zhang. A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer. *Scientific Data*, 10(1):231, 2023.
- [59] Jevgenij Gamper, Navid Alemi Koohbanani, Simon Graham, Mostafa Jahanifar, Ksenija Benet, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. Pannuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778*, 2020.
- [60] Patrick Wagner, Maximilian Springenberg, Marius Kröger, Rose KC Moritz, Johannes Schleusener, Martina C Meinke, and Jackie Ma. Semantic modeling of cell damage prediction: a machine learning approach at human-level performance in dermatology. *Scientific Reports*, 13(1):8336, 2023.
- [61] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihhan Simsek, Seval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *arXiv preprint arXiv:2403.17834*, 2024.
- [62] Jun Ma and Bo Wang. Miccai flare23: Fast, low-resource, and accurate organ and pan-cancer segmentation in abdomen ct, Apr 2023.
- [63] Amir Akbarnejad, Nilanjan Ray, Penny J Barnes, and Gilbert Bigras. Predicting ki67, er, pr, and her2 statuses from h&e-stained breast cancer images. *arXiv preprint arXiv:2308.01982*, 2023.
- [64] Pengfei Shao, Lijun Tang, Pu Li, Yi Xu, Chao Qin, Qiang Cao, Xiaobing Ju, Xiaoxin Meng, Qiang Lv, Jie Li, Wei Zhang, and Changjun Yin. Precise segmental renal artery clamping under the guidance of dual-source computed tomography angiography during laparoscopic partial nephrectomy. *Eur. Urol.*, 62(6):1001–1008, December 2012.
- [65] Pengfei Shao, Chao Qin, Changjun Yin, Xiaoxin Meng, Xiaobing Ju, Jie Li, Qiang Lv, Wei Zhang, and Zhengquan Xu. Laparoscopic partial nephrectomy with segmental renal artery clamping: technique and clinical outcomes. *Eur. Urol.*, 59(5):849–855, May 2011.
- [66] Y He, G Yang, J Yang, Y Chen, Y Kong, J Wu, L Tang, X Zhu, JL Dillenseger, P Shao, S Zhang, H Shu, JL Coatrieux, and S Li. Dense biased networks with deep priori anatomy and hard region adaptation: Semisupervised learning for fine renal artery segmentation. *Medical Image Analysis*, 63, 2020.
- [67] Yuting He, Guanyu Yang, Jian Yang, Rongjun Ge, Youyong Kong, Xiaomei Zhu, Shaobo Zhang, Pengfei Shao, Huazhong Shu, Jean-Louis Dillenseger, Jean-Louis Coatrieux, and Shuo Li. Meta grayscale adaptive network for 3D integrated renal structures segmentation. *Med. Image Anal.*, 71:102055, July 2021.
- [68] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.
- [69] Meng Lou, Hanning Ying, Xiaoqing Liu, Hong-Yu Zhou, Yuqing Zhang, and Yizhou Yu. Sdr-former: A siamese dual-resolution transformer for liver lesion classification using 3d multi-phase imaging. *arXiv preprint arXiv:2402.17246*, 2024.
- [70] Lidia Garrucho, Claire-Anne Reidel, Kaiser Kushibar, Smriti Joshi, Richard Osuala, Apostolia Tsirikoglou, Maciej Bobowicz, Javier del Riego, Alessandro Catanese, Katarzyna Gwoździewicz, et al. Mama-mia: A large-scale multi-center breast cancer dce-mri benchmark dataset with expert segmentations. *arXiv preprint arXiv:2406.13844*, 2024.

- [71] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.
- [72] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.
- [73] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-ray: Hospital-scale chest x-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. In *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*, Advances in computer vision and pattern recognition, pages 369–392. Springer International Publishing, Cham, 2019.
- [74] Masayuki Tsuneki and Fahdi Kanavati. Inference of captions from histopathological patches. In *International Conference on Medical Imaging with Deep Learning*, pages 1235–1250. PMLR, 2022.
- [75] Masakata Kawai, Noriaki Ota, and Shinsuke Yamaoka. Large-scale pretraining on pathological images for fine-tuning of small pathological benchmarks. In *Workshop on Medical Image Learning with Limited and Noisy Data*, pages 257–267. Springer, 2023.
- [76] Yuri Tolkach, Lisa Marie Wolgast, Alexander Damanakis, Alexey Pryalukhin, Simon Schallenberg, Wolfgang Hulla, Marie-Lisa Eich, Wolfgang Schroeder, Anirban Mukhopadhyay, Moritz Fuchs, et al. Artificial intelligence for tumour tissue detection and histological regression grading in oesophageal adenocarcinomas: a retrospective algorithm development and validation study. *The Lancet Digital Health*, 5(5):e265–e275, 2023.
- [77] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.
- [78] Carsen Stringer and Marius Pachitariu. Cellpose3: one-click image restoration for improved cellular segmentation. *bioRxiv*, pages 2024–02, 2024.
- [79] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.
- [80] N. Gaggion, C. Mosquera, M. Aineseder, L. Mansilla, D. Milone, and E. Ferrante. CheXmask Database: a large-scale dataset of anatomical segmentation masks for chest x-ray images (version 0.1). <https://doi.org/10.13026/dx54-8351>, 2023.

Appendix

Supplementary material

We present the following items in the supplementary material section:

1. Data source about MedTrinity-25M. (Section A)
2. Quantitative comparison between GPT-4V and LLaVA-Med Captioner (Section B).
3. Example of ROI for normal regions (Section C).
4. The list of expert ROI models (Section D).
5. Prompt for evaluating MedTrinity-25M alignment with human annotations (Section E).
6. Prompt for generating MedTrinity-25M. (Section F).
7. A Datasheet [52] for MedTrinity-25M (Section G).

A Data Source

Table 4: Data sources for MedTrinity-25M from various medical image datasets, detailing their modalities, biological structures, quantities, and annotations.

Dataset Name	Modality	Biological Structures	Quantity	Text	Disease Type	BBox	Mask
BHX[53]	MRI	brain	973908	✗	✗	✗	✓
BRATS24-MICCAI[54]	MRI	brain	2535132	✗	✗	✓	✗
BRATS-ISBI[55]	MRI	brain	987340	✗	✗	✓	✗
breast histopathology[56, 57]	Histopathology	breast	547403	✗	✓	✗	✗
BreastCancer[58]	Histopathology	breast	1824	✗	✗	✓	✗
CheXpert[7]	X-Ray	lung	183242	✗	✓	✗	✗
CISC[59]	Histopathology	Adrenal, Bile duct, Bladder, Breast, Colon, Cervix, Esophagus Kidney, Liver,etc	16285	✗	✓	✓	✗
CPD[60]	Histopathology	skin	204	✗	✗	✓	✗
CT-RATE[61]	CT	lung, liver, mediastinum, kidney, heart, etc.	3869640	✓	✗	✗	✗
DeepLesion[40]	CT	bone, abdomen, mediastinum, liver, lung, kidney, soft tissue, pelvis	2889672	✗	✗	✗	✓

Table 4 : Continued from previous page

Dataset Name	Modality	Biological Structures	Quantity	Text	Disease Type	BBox	Mask
FLARE23[62]	CT	Liver, kidney, spleen, pancreas, Aorta, adrenal gland, Gallbladder, esophagus, stomach, duodenum,etc.	13770	✗	✓	✓	✗
ihc4bc[63]	Microscopy	cell	102535	✗	✓	✗	✗
KIPA22[64, 65, 66, 67]	CT	kidney, cervix	26878	✗	✗	✓	✗
LLaVA-Med[68]	CT, MR, Endoscopy, X-Ray, Ultrasound, Histopathology, Dermoscopy, Microscopy, Fundus, PET	cell, rib, tissue, face, brain, vascular, liver, bone, lymph, etc.	22550	✓	✗	✗	✗
LLD-MMRI[69]	MRI	liver	21523	✗	✗	✓	✗
MAMA-MIA[70]	MRI	breast	316113	✗	✗	✓	✗
MIMIC-CXR-JPG[8]	X-Ray	lung	240506	✓	✓	✗	✓
NCT-CRC-HE-100K[43]	Histopathology	colon	100361	✗	✓	✗	✗
NIH-CXR[71, 72, 73]	X-Ray	lung	986	✗	✗	✗	✓
PadChest[6]	CT	lung	96284	✓	✗	✗	✗
PatchGastricADC22[74]	MRI	brain	98399	✗	✓	✗	✗
Path-VQA training[45]	Pathology	gastrointestinal, colon, appendix, pinworm,etc.	13375	✓	✓	✗	✗
PMC-OA[24]	CT, MR, Endoscopy, X-Ray, Ultrasound, Histopathology, Dermoscopy, Microscopy, Fundus, PET	cell, tissue, vascular, brain, bone, liver, lymph, eye, epithelium, etc.	856999	✓	✗	✗	✗
PMC-VQA[28]	CT, MR, Endoscopy, X-Ray, Ultrasound, Histopathology, Dermoscopy, Microscopy, Fundus, PET	cell, brain, tissue, artery, bone, face, rib, vascular, liver, eye, etc.	144999	✓	✗	✗	✗
PTCGA[75]	Histopathology	brain, breast, uterine corpus, kidney, lung, thyroid	3293965	✗	✓	✓	✗

Table 4 : Continued from previous page

Dataset Name	Modality	Biological Structures	Quantity	Text	Disease Type	BBox	Mask
Quilt-1M[10]	Histopathology	skin, lung, soft tissue, blood, kidney, bone, etc.	643819	✓	✗	✗	✗
SAMMed-20M[44]	X-Ray, PET, CT, MR, Endoscopy, dermoscopy	brain, kidney, liver, lung, pancreas, pulmonary, hepatic, skin, etc.	5491274	✗	✓	✓	✗
SLAKE training[22]	CT, MRI, X-Ray	brain, liver, kidney, pelvic, lung	646	✓	✓	✓	✗
TCGA[75]	Histopathology	tissue	1142221	✗	✗	✓	✗
ULS23	CT	lung, lymph nodes, bladder, brain, colon, kidney, lung, pancreas.	105669	✗	✗	✓	✗
VALSET[76]	Histopathology	oesophagus, stomach	277565	✗	✓	✗	✗
VQA-RAD training[77]	X-Ray, MRI	brain, lung, abdomen, etc.	1758	✓	✓	✗	✗
Total			25016845				

B Quantitative Comparison of LLaVA-Med++ with GPT-4V

As detailed in Section 3.2.2 of the main paper, we developed an enhanced version of LLaVA-Med [9], called LLaVA-Med++. This enhancement leverages the latest LLaMA3 [36] to boost linguistic capabilities and incorporates multi-scale feature extraction [37] to improve vision capabilities.

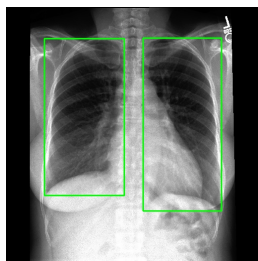
To justify the selection of our specialized medical model, LLaVA-Med++, over GPT-4V for generating textual descriptions, we conducted a quantitative comparison of the outputs generated by both models. We assessed the level of detail by comparing the average word count of text descriptions generated for the same sample. As shown in Figure 12, LLaVA-Med++, after task-specific fine-tuning, outperformed GPT-4V by 3.6% in word count, indicating that the descriptions generated by LLaVA-Med++ are more detailed. Based on these findings, we selected LLaVA-Med++ to generate fine-grained captions for our entire MedTrinity-25M.

Figure 12: Qualitative comparison of the relative average word count of samples generated by LLaVA-Med++ and GPT-4V.



Figure 13: Examples of ROIs for normal regions.

(a) A no infection sample from MIMIC-CXR. The ROIs highlight the left and right lungs.



(b) A healthy sample from SLAKE. The ROI points out the liver.

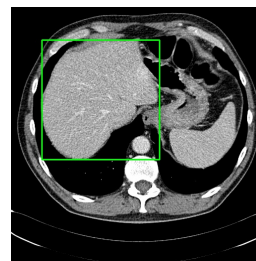


Table 5: List of expert models used to generate ROIs for different datasets.

ID	Dataset Name	Model
1	breast histopathology	HoverNet [78]
2	BreastCancer	
3	CISC	
4	CPD	
5	NCT-CRC-HE-100K	
6	PTCGA	
7	TCGA	
8	VALSET	
9	ihc4bc	
10	Quilt-1M	
11	CT-RATE	SAT [29]
12	PMC-OA	DINO [79]
13	PMC-VQA	
14	LLaVA-Med	
15	Path-VQA training	
16	PadChest	CheXmask [80] [31]
17	MIMIC-CXR-JPG	
18	CheXpert	

C Examples of ROIs for Normal Regions

As detailed in Section 3.1 of the main paper, the regions of interest (ROIs) identified using expert grounding models predominantly contain pathological findings such as lesions, inflammation, neoplasms, infections, or other potential abnormalities. In the few instances where no abnormalities are present, the ROIs typically highlight the primary object or organ in the image. Examples of ROIs without abnormalities are shown in Figure 13.

D List of Expert models to locate ROIs

As detailed in Section 3.2.1 of the main paper, for datasets lacking localization information such as segmentation masks and bounding boxes, we employ various pretrained expert models to identify the ROIs. The specific expert models used for each dataset are listed in Table 5.

E Evaluation Prompt of Alignment to Human Annotations

The prompt used to query GPT-4V for evaluating the alignment score is shown in Figure 14.

Figure 14: Prompt used to evaluate the alignment of generated fine-grained captions.

Prompting MLLMs to evaluate the alignment of generated multi-granular annotations with human annotations

Let's think it step by step. Evaluate the multigranular radiology report annotations (Report A) compared to the radiology report B step by step. Both reports are based on the same image. Follow these guidelines to ensure accurate assessment:

Note: If neither the original question nor radiology report B mentions any abnormalities or diseases, such as "the lungs are clear without confluent consolidation or effusion" or "no pneumothorax is seen", skip the evaluation and return "None."

Basic Rating Rules:

1. Evaluate each attribute in Report A against radiology report B and verify the information by analyzing the image. Do not deduct points without image analysis.
2. Judge correctness based on the accuracy of medical facts and diagnoses, not on the exact alignment of sentence structure or organization.
3. If radiology report B does not mention any abnormalities or diseases, skip the evaluation and return "None," such as "the lungs are clear without confluent consolidation or effusion" or "no pneumothorax is seen".
4. Each of the 5 attributes should be judged independently. Errors in one attribute should not affect the scoring of other attributes.

Attributes and Corresponding Rating Rules:

1. **Modality Used for Imaging:**
 - **Rating Rule:** Compare with radiology report B. Different names for the same modality (e.g., "chest X-ray" and "CXR") are acceptable.
2. **Specify the Organ and Anatomical Structures:**
 - **Rating Rule:** Check if the organs and anatomical structures in Report A match those in radiology report B or appear in the image.
 - Mentioned in both: 2 points
 - Mentioned in one: 1 point
 - Not mentioned in either: 0 points
 - Do not deduct points without image analysis.
3. **Locations of ROI (Regions of Interest):**
 - **Rating Rule:** Compare the "horizontal" and "vertical" positions, and the "area ratio" of ROIs with radiology report B. A 5% error in the area ratio is acceptable. If Report A includes at least one ROI from radiology report B, no points are deducted, even if all ROIs are not covered.
4. **Analysis of Abnormal Characteristics:**
 - **Rating Rule:** Characteristics indicating pathology should match those in radiology report B or appear in the image.
 - Mentioned in both: 2 points
 - Mentioned in one: 1 point
 - Not mentioned in either: 0 points
 - Do not deduct points without image analysis.
5. **Comparison of Lesions and Surrounding Regions:**
 - **Rating Rule:** Differences in features and disease progression should match those in radiology report B or appear in the image.
 - Mentioned in both: 2 points
 - Mentioned in one: 1 point
 - Not mentioned in either: 0 points
 - Do not deduct points without image analysis.

Note: Return the scores in a list. For example, if attributes 4 and 5 get deducted 1 point each, while others score 2 points each, return [2, 2, 2, 1, 1]. Provide a short reason (within 80 words) for each point deduction.

Prompting MLLMs to generate multigranular textual description

```
caption_template = Template("<image>
`Caption of the image` : {{caption}}
`Disease or organ` : {{disease}}
`Specific position` : {{descs}}
`Knowledge` : {{knowledge}}
You are provided with a biomedical image from a medical dataset, the disease type (or organ name if there is no disease) of the dataset ('Disease or organ'), the medical Knowledge of the disease ('Knowledge') and a coarse caption ('Caption') of the image. In addition, the green bounding box and its specific position in the image ('Specific position') are given, indicating appearance of disease. If no green bounding box, there is no disease.
Your task is to answer the following questions based on the image, green bounding box, caption, disease type and disease knowledge, and condense your answers into caption-styled text.
### question1
Give me a detailed description of the image, including type of the image, organs in the image, approximate location of these organs and relevant locations of these organs and any medical devices (if present) visible in the image as detailedly as possible.
Note when answering question1:
1. Not all disease knowledge is relevant to this image; only utilize disease knowledge pertinent to the condition depicted in this image for analysis.
2. The coarse caption may not explicitly describe the image, for example, there may appear multiple organs in the caption. You should utilize your knowledge to figure out the most ONE organ and ONE disease to give your description.
3. Your answer should not contain anything about the green bounding box like the contour itself and its outline.
4. Do not explain or emphasize your analysis.
### question2
Specify the specific location of the green bounding box in the image and its relative position to other reference objects in the image. Describe what is unusual in the green bounding box indicating the disease (color, texture, size and other features) .
Note when answering question2:
1. "specific location" is the given parameter `Specific position` but "relative position" is not provided.
2. There may be multiple green bounding boxes, and the contents of these contours may not necessarily represent the affected areas. Therefore, you need to first answer the questions based on the contents within each green bounding box. Afterward, analyze the location of the disease based on your answers.
3. Do not use phrase "green bounding box" in your response, use "region of interest" as a substitution. Do not contain phrases "caption", "medical annotation", "medical knowledge".
4. Do not say anything that is not needed in your analysis, like introduction of the disease and medical equipments.
5. Do not explain or emphasize your analysis.
### question3
What may be the relationship between the content in the green bounding box and other regions (others being cause of the disease/jointly affected by the diseases/one affect the others/relative positional relationships)? Why and is it possible?
Note when answering question3:
1. Utilize external knowledge, if possible, to choose relationships and give necessary analysis.
2. You can only give an explanation to your choice within two sentences.
3. Do not summarize what you've said.
4. Do not emphasize your analysis.
### Integrate Information
Describe your answers in a descriptive sentence, not in a "Question-Answer" style. Combine and slightly shorten your answers to the above three questions into a coherent text, keeping as much information of your answers as possible.
Note when integrating information and outputting your response:
1. Don't respond saying you're unable to assist with requests.
2. You should only output your combined and shortened text.
")
prompt = caption_template.render([caption, disease, knowledge, loc_descs])
```

F Prompt Template for Generation of Multigranular Text Description

To generate multigranular textual descriptions, we design a multi-task prompting approach, breaking down this task into several smaller descriptive tasks. The model’s responses to these different tasks collectively form the final fine-grained text description.

appendix F illustrates our prompt template consisting of a three-level hierarchical framework with questions to instruct MLLMs:

Step 1 - Global Understanding: Instruct MLLMs to provide a comprehensive description of the image, detailing all modalities, identified anatomical structures, and their approximate locations. This step ensures that MLLMs gains an overarching understanding and basic information about the image.

Step 2 - Local Analysis: Instruct MLLMs to conduct a detailed analysis of the regions of interest (ROI), including their locations, abnormalities, and textures. This step guides MLLMs to focus on specific lesions for a thorough assessment.

Step 3 - Local-Global Relationship: Instruct MLLMs to examine the relationship between local and global regions and predict how the surrounding areas will be affected by the lesions in the ROI. This step aims to understand the interaction between local and global attributes, assessing the impact of local abnormalities on the entire organ for accurate disease diagnosis.

G Datasheet for MedTrinity-25M

In this section, we present a DataSheet [52] for MedTrinity-25M, synthesizing many of the other analyses we performed in this paper.

1. Motivation For Datasheet Creation

- **Why was the dataset created?** The dataset was created to provide a large-scale, multimodal, multigranular medical dataset to support a wide range of multimodal tasks such as captioning, report generation, classification, and segmentation. It aims to facilitate large-scale pre-training of multimodal medical AI models by providing enriched annotations from unpaired image inputs.
- **Has the dataset been used already?** Yes. Multigranular annotations enable a wide range of tasks like Medical Visual Question Answering, which we discuss in Section 4.
- **What (other) tasks could the dataset be used for?** The MedTrinity-25M dataset could be used for multiple medical imaging tasks such as classification, segmentation, detection, and medical report generation. Its extensive and detailed annotations make it suitable for training and evaluating machine learning models across these tasks.
- **Who funded dataset creation?** This work is partially supported by the OpenAI Researcher Access Program, AWS Cloud Credit for Research Program, TPU Research Cloud (TRC) program and Google Cloud Research Credits program.

2. Data composition

- **What are the instances?** Each instance in the dataset is a triplet consisting of an image, a Region of Interest (ROI), and a multigranular textual description. The ROI is associated with abnormalities and represented by bounding boxes or segmentation masks.
- **How many instances are there?** The dataset comprises over 25 million image-ROI-description triplets sourced from more than 90 online resources, spanning 10 modalities and covering over 65 diseases.
- **What data does each instance consist of?** Each instance consists of a medical image, a corresponding ROI (highlighting abnormalities within the image), and a detailed, multigranular textual description that includes disease/lesion type, modality, region-specific description, and inter-regional relationships.
- **Is there a label or target associated with each instance?** Yes, the textual description serves as a detailed label or target, providing information about the disease or lesion type, as well as other relevant medical details.
- **Is any information missing from individual instances?** No.
- **Are relationships between individual instances made explicit?** Not applicable – we do not study relationships between disparate medical samples.
- **Does the dataset contain all possible instances or is it a sample?**
Our generation pipeline includes all instances collected from available medical data sources. However, the current list of medical dataset sources is not exhaustive, indicating a high probability of collecting additional instances in the future.

- **Are there recommended data splits (e.g., training, development/validation, testing)?** There are no recommended data splits, as this data was curated mainly for pretraining rather than evaluation.
- **Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.** Yes. Despite multiple efforts to minimize errors using coarse captions and external medical knowledge, the textual descriptions generated by MLLMs may still contain inaccuracies.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** The dataset is largely self-contained. However, it was constructed using data from over 90 online resources such as TCIA, Kaggle, Zenodo, and Synapse. The images and related data were collected from these sources, but the dataset itself does not rely on external resources like websites or tweets for its primary functionality once compiled.

3. Collection Process

- **What mechanisms or procedures were used to collect the data?** The data collection involved an automated pipeline that scales up multimodal data by generating multigranular visual and textual annotations from unpaired images. Data was collected from over 90 different sources, preprocessed, and grounded using domain-specific expert models to identify ROIs related to abnormal regions.
- **How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data?**
The data associated with each instance was indirectly inferred and derived from the collected images using domain-specific expert models and multimodal large language models (MLLMs). The images were annotated with bounding boxes, segmentation masks, and textual descriptions, transforming them into image-ROI-description triplets.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?** The dataset is not a sample from a larger set but an extensive collection aggregated from multiple datasets and online sources. The strategy was to include as many diverse images and annotations as possible from a wide range of medical datasets.
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?** Data collection was primarily done by the co-authors of this paper.
- **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.** The data was collected from April 2024 to June 2024.

4. Data Preprocessing

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** Extensive preprocessing and annotation were performed, including segmentation, bounding box creation, and generating multigranular textual descriptions. The preprocessing also involved integrating metadata and knowledge retrieval from sources like PubMed to create comprehensive descriptions.
- **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the ‘raw’ data.** The raw data was saved, but at this time we do not plan to release it directly due to copyright and privacy concerns.
- **Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.** The software for preprocessing and labeling, including the automated pipeline and MLLMs, is available at <https://github.com/yunfeixie233/DataProcessingSystem>.
- **Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?** Yes. The preprocessing and collection procedures align with the motivation of creating a comprehensive, large-scale multimodal dataset to support the development of advanced medical AI models. The dataset’s multigranular annotations enable a wide range of tasks like Medical Visual Question Answering, which we discuss in Section 4.

5. Dataset Distribution

- **How will the dataset be distributed?** The dataset is publicly available and can be accessed via the provided link: MedTrinity-25M <https://yunfeixie233.github.io/MedTrinity-25M/>.
- **When will the dataset be released/first distributed? What license (if any) is it distributed under?** We will release it as soon as possible, using a permissible license for research-based use.
- **Are there any copyrights on the data?** We believe our use is ‘fair use,’ however, due to an abundance of caution, we will not be releasing any of the videos themselves.
- **Are there any fees or access restrictions?** No.

6. Dataset Maintenance

- **Who is supporting/hosting/maintaining the dataset?** The first authors of this paper.
- **Will the dataset be updated? If so, how often and by whom?** We do not plan to update it at this time.
- **Is there a repository to link to any/all papers/systems that use this dataset?** Not right now, but we encourage anyone who uses the dataset to cite our paper so it can be easily found.
- **If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?** Not at this time.

7. Legal and Ethical Considerations

- **Were any ethical review processes conducted (e.g., by an institutional review board)?** No official processes were done, as our research is not on human subjects, however, because the dataset is in the medical domain we had significant internal discussions and deliberations when choosing the scraping strategy.
- **Does the dataset contain data that might be considered confidential?** The dataset does not contain data that might be considered confidential, as it uses publicly available sources and anonymized medical data.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why?** The dataset does not contain data that might be offensive, insulting, threatening, or anxiety-inducing. It consists of medical images and associated annotations for clinical and research use.
- **Does the dataset relate to people?** The dataset relates to people as it involves medical images and data. However, it is anonymized and does not include identifiable information.
- **Does the dataset identify any subpopulations (e.g., by age, gender)?** Not explicitly (e.g. through labels)
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** The dataset does not identify specific subpopulations directly in the provided description. Additionally, it is not possible to identify individuals from the dataset as it is anonymized and compiled from various sources.