Uncertainty-aware Medical Diagnostic Phrase Identification and Grounding

Ke Zou, Yang Bai, Bo Liu, Yidi Chen, Zhihao Chen, Yang Zhou, Xuedong Yuan, Meng Wang, Xiaojing Shen, Xiaochun Cao *Senior Member, IEEE*, Yih Chung Tham, Huazhu Fu *Senior Member, IEEE*

***Abstract*—Medical phrase grounding is crucial for identifying relevant regions in medical images based on phrase queries, facilitating accurate image analysis and diagnosis. However, current methods rely on manual extraction of key phrases from medical reports, reducing efficiency and increasing the workload for clinicians. Additionally, the lack of model confidence estimation limits clinical trust and usability. In this paper, we introduce a novel task called Medical Report Grounding (MRG), which aims to directly identify diagnostic phrases and their corresponding grounding boxes from medical reports in an end-to-end manner. To address this challenge, we propose uMedGround, a robust and reliable framework that leverages a multimodal large language model to predict diagnostic phrases by embedding a unique token, <BOX>, into the vocabulary to enhance detection capabilities. A vision encoder-decoder processes the embedded token and input image to generate grounding boxes. Critically, uMedGround incorporates an uncertainty-aware prediction model, significantly improving the robustness and reliability of grounding predictions. Experimental results demonstrate that uMedGround outperforms state-of-the-art medical phrase grounding methods and fine-tuned large visual-language models, validating its effectiveness and reliability. This study represents a pioneering exploration of the MRG task, marking the first-ever endeavor in this domain. Additionally, we demonstrate the applicability of uMedGround in medical visual question answering and class-based localization tasks, where it highlights visual evidence aligned with key diagnostic phrases, supporting clinicians in interpreting various types of textual inputs, including free-text reports, visual question answering queries, and class labels.**

***Index Terms*—Medical report grounding, vision-language model, uncertainty estimation.**

Manuscript received December, 2024; revised June, 2025. This work was supported by the National Key Researchand Development Program of China under Grant 2020YFA0714003, the National Natural Science Foundation of China (No.62025604, 62411540034), the H. Fu's Agency for Science, Technology and Research (A*STAR) Central Research Fund ("Robust and Trustworthy AI system for Multi-modality Healthcare"), the Science and Technology Department of Sichuan Province (Grant No. 2022YFS0071), the Science and Technology Department of Guangxi Zhuang Autonomous Region (Grant No. 2025GXNSFAA069531), Science and Technology Department of Hainan Province (Grant number ZDYF2024SHFZ052) and the China Scholarship Council (No. 202206240082).

K. Zou and X. Yuan are with the College of Computer Science, Sichuan University, Chengdu, China.

Y. Bai, Y. Zhou, M. Wang, and H. Fu are with the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore.

Bo Liu is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China.

Y. Chen is with the Department of Radiology, West China Hospital, Sichuan University, Chengdu, China.

Z. Chen is with the College of Intelligence and Computing, Tianjin University, Tianjin, China

X. Shen is with the Department of Mathematics, Sichuan University, Chengdu, China.

X. Cao is with the School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-Sen University, Shenzhen, China.

K. Zou and Y. Tham are with the Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore and the Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore.

Ke Zou and Yang Bai are equally contributed. Xuedong Yuan and Huazhu Fu are the co-corresponding authors (e-mail: yxdongdong@163.com, hzfu@ieee.org).

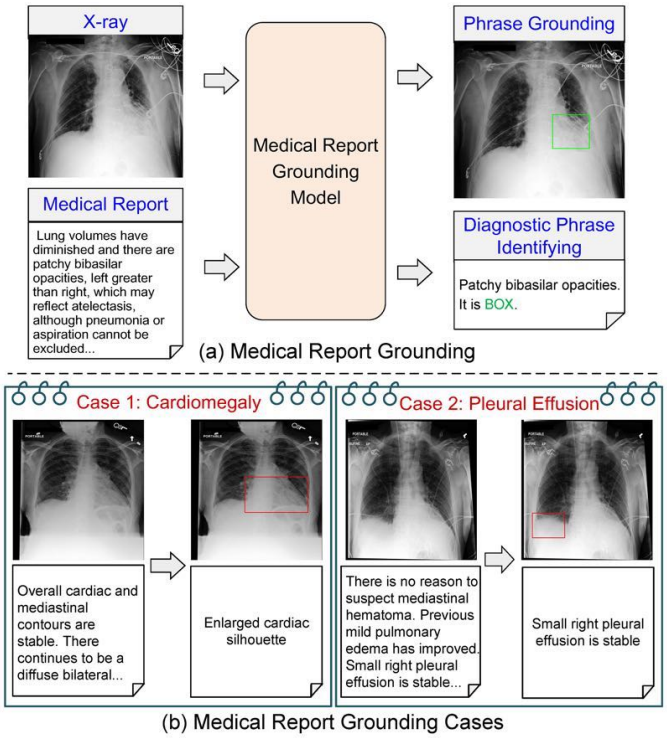

Fig. 1. Medical report grounding. (a) Illustration of our medical report grounding with large-language model for radiological diagnosis. (b) Different paired cases (X-ray image, medical report, grounding box and key phrase): Cardiomegaly and Pleural Effusion.

I. Introduction

Medical image grounding involves detecting bounding boxes of corresponding categories from medical images, assisting clinicians in identifying abnormalities. To facilitate a more direct approach for localizing areas that clinicians need to sketch based on diagnostic phrases, Chen et al. [1] introduced the medical phrase grounding. However, this task requires first extracting the relevant phrase from a medical report and then detecting the corresponding region in the image. In this paper, we propose a novel task, **M**edical **R**eport **G**rounding (**MRG**), which directly identifies diagnostic phrase and their

corresponding grounding box from the medical report. As illustrated in Fig. 1 (a), this task presents two key challenges: identifying diagnostic phrase within the report and grounding bounding box associated with it. Fig. 1 (b) provides additional examples using X-ray imaging, showcasing cases such as cardiac hypertrophy and pleural effusion diagnoses. This process plays a crucial role in enhancing the comprehension of semantic information within medical reports, enabling the identification of grounding box associated with phrase in medical image.

In computer vision, tasks similar to MRG include visual grounding, which has been extensively explored in the context of natural images [2–6]. In contrast, its application to medical imaging remains limited, primarily due to challenges in labeling and acquiring medical data. Chen et al. [1] and Ichinose et al [7] first introduced methods for medical phrase X-ray grounding and medical report CT grounding. However, their approaches rely heavily on pre-processing steps to extract diagnostic phrases from medical reports, which limits system efficiency and scalability. Recent advances in large vision-language models (VLMs), such as GPT-4V [8], LLaVA-13B [9], and InternVL [10, 11], have shown strong capabilities in multi-modal understanding and reasoning. However, the precision and reliability of these models in identifying diagnostic phrase and grounding corresponding bounding box remain to be fully optimized. These models combine large language models (LLMs) with vision encoders to directly generate textual responses from input images. While promising, their ability to precisely identify diagnostic phrases and localize corresponding regions with bounding boxes remains suboptimal, especially in clinical deployment scenarios where interpretability and reliability are paramount.

To this end, an important but often overlooked aspect is the uncertainty associated with model predictions, especially bounding box predictions in medical contexts. Unlike natural image tasks, vague or overly confident localization in clinical settings could lead to critical misinterpretations. Uncertainty estimation offers a potential solution by transforming single-point predictions into probabilistic or multi-hypothesis representations. Common uncertainty estimation methods in medical domain include Monte-Carlo Dropout (MCDO) [12], Ensemble [13], Test-time data augmentation (TTA) [14], evidential-based [15–17] and deterministic-based methods [18–20]. However, such techniques have rarely been adapted to phrase grounding, and direct integration into large VLMs may incur prohibitive computational costs. To address these limitations, we propose a lightweight and scalable multi-hypothesis prediction framework, which serves as a plug-and-play module within vision-language models. This design improves grounding robustness without significantly increasing inference time or model complexity. Importantly, it provides variance-aware outputs, offering interpretable visual cues that can enhance trust and adoption in clinical workflows.

In this paper, we introduce a new task, medical report grounding, which addresses two key challenges: identifying diagnostic phrase and grounding their corresponding box. To tackle these challenges, we propose an innovative framework **u**ncertainty-aware **Me**dical **d**iagnostic phrase identification and **ground**ing, named **uMedGround**. It begins by leveraging a multimodal LLM to enhance the understanding and interpretation of medical report, enabling accurate identification of diagnostic phrase. Once the multimodal LLM identifies the diagnostic phrase, a novel token, $<$ BOX $>$ is embedded into the vocabulary to represent the phrase prediction. Then, the implicit embedding of $<$ BOX $>$, together with the input medical image, is then decoded using a SAM-based encoder-decoder to generate the corresponding grounding box. Most critically, we introduce an uncertainty-aware prediction model to quantify confidence in the grounding box, enhancing both robustness and reliability. This is achieved by integrating a multi-hypothesis prediction head into the fine-tuning process of the box decoder, enabling the simultaneous output of the box and its associated uncertainty. Our contribution can be summarized as follows

- We propose a novel task, medical report grounding, which involves diagnostic phrase identification and grounding box prediction. This task assists doctors in identifying key diagnostic information and its visual cues within medical reports and images, facilitating improved diagnostic accuracy and supporting more informed clinical decision-making.
- An end-to-end framework is developed in this study, uMedGround, which incorporates a reliable multimodal LLM to enhance the understanding and interpretation required for medical report grounding.
- A new <BOX> token is introduced as an embedding input to the box decoder, significantly improving the precision of critical finding localization within medical images.
- We pioneer the integration of uncertainty estimation into medical report grounding, enhancing both the accuracy and reliability of grounding predictions.
- Three publicly available datasets are preprocessed and curated to generate four paired subdatasets: two designed for experimental validation and two intended for clinical application testing. Comprehensive experiments and applications on these datasets demonstrate the accuracy, reliability, and applicability of the proposed framework. This study represents a pioneering exploration of the MRG task, marking the first-ever endeavor in this domain.[1]

II. Related works

A. Visual-language model for grounding

Visual grounding had been extensively studied in the context of natural images, evolving from early two-stage models [21, 22] to multi-modal transformer-based approaches [2–6]. However, these methods faced challenges when applied to radiological images. MRG demanded specialized visual-textual feature learning to enable the model to discern medical findings that exhibited subtle textural and morphological variations and to interpret the spatial relations described in medical reports. The transition from general visual grounding to the more specialized domain of MRG underscored the need for tailored approaches.

[1]Our codes and curated datasets will be released in https://github.com/Cocofeat/uMedGround.

This need was reflected in several pivotal studies. Moon et al. [23] leveraged large-scale medical image-text data for vision-language pretraining, highlighting the potential of AI in understanding complex multimodal medical narratives. Wang et al. [24] proposed a cross-modal contrastive attention model that bridged the gap between radiology images and reports through aligned multi-modal representations. Despite these advancements in knowledge representation, a performance gap remained in translating this understanding into accurate and clinically valid grounding predictions. More recently, large vision-language models (VLMs) such as GPT-4V [8], LLaVA-13B [9], and InternVL [10, 11] demonstrated significant potential for multimodal understanding and reasoning. In the medical domain, visual-language pretraining frameworks like MedKLIP [25] and BioViL [26] incorporated clinical knowledge to improve grounding performance, enabling fine-grained contextual understanding critical for downstream tasks. Among early MRG efforts, MedRPG [1] adopted a transformer-based model to predict bounding boxes given pre-extracted medical phrases, while Ichinose et al. [7] introduced a framework for grounding in 3D CT-report pairs by structuring reports to guide phrase localization. Despite their effectiveness, these methods required explicit report preprocessing to isolate key phrases, which limited their scalability and adaptability.

To overcome these limitations, we proposed a method that directly extracted key phrases from free-text reports using LLMs, allowing for intuitive understanding and seamless end-to-end grounding. In a similar direction, Vilouras et al. [27] employed a generative foundation model for zero-shot medical phrase grounding and achieved competitive results. While large-scale VLMs for medicine remained underexplored, existing work primarily targeted tasks such as disease classification, visual question answering (VQA), and radiology report generation [28–30]. To address the long-tail distribution of clinical knowledge and enhance diagnostic reasoning, recent studies integrated heterogeneous external sources, including multi-source retrieval [31] and medical knowledge graphs [32], into LLM-based QA systems. MARIA-2 [33] advanced radiology report generation by evaluating accuracy and completeness, although it remained task-specific. In contrast, our model aimed to deliver a unified and generalizable solution for grounding that could also be extended to classification-based localization and VQA tasks.

B. Uncertainty estimation in medical image analysis

Uncertainty estimation had become an increasingly important topic in medical image analysis, serving as a means for improving reliability and interpretability. According to Huang et al. [34], uncertainty estimation had been explored through both probabilistic [12, 14, 35–37] and non-probabilistic [16, 17, 38] approaches across various medical imaging tasks.

For probabilistic methods, Kohl et al. [35] introduced a probabilistic U-Net for segmenting ambiguous medical images, though its computational cost was prohibitively high. Later, Nair et al. [12] investigated MC dropout and a range of uncertainty metrics in deep learning for multiple sclerosis lesion segmentation. Ensemble learning also emerged as a viable approach; Mehrtash et al. [36] used ensembles to improve confidence calibration and predictive reliability in deep segmentation networks.

On the non-probabilistic side, Wang et al. [14] proposed a brain tumor segmentation method that utilized test-time augmentation to estimate uncertainty and boost robustness. However, these techniques typically required multiple inferences per image and thus introduced additional computational burdens. To address this limitation, Zou et al. [16] and Wang et al. [17] applied evidential deep learning to brain tumor and retinal disease segmentation, offering single-pass uncertainty estimation that improved both robustness and generalizability. In summary, although uncertainty estimation methods have seen significant application in medical classification and segmentation, their integration into medical grounding tasks is still limited. Furthermore, directly incorporating these methods into medical grounding tasks may significantly increase model parameters and inference time, particularly for VLMs.

III. Methodology

A. Problem Definition

Medical report grounding aims to extract crucial diagnostic information, resembling the expertise of an X-ray specialist, from an original medical report $\mathbf{x}_t$ and an X-ray image $\mathbf{x}_i$. Specifically, the objective is to identify the medical phrase $\mathbf{y}_p$ and its corresponding detection grounding box $\mathbf{y}_b$ within the X-ray image. This process can be summarized as follows:

$$(\mathbf{y}_p, \mathbf{y}_b) = \mathcal{F}(\mathbf{x}_i, \mathbf{x}_t), \tag{1}$$

$b = (x, y, w, h)$, (x, y) denote the center coordinates of the bounding box in the image, while (w, h) correspond to its height and width, respectively. Furthermore, $\mathcal{F}(\cdot)$ represents the mapping function that relates the input variables to the output. While Chen et al. [1] and Ichinose et al [7] proposed medical report grounding, their approaches necessitate manual preprocessing of the phrases. Leveraging the reasoning capabilities of large VLMs, our method automatically infers phrases from original medical reports and conducts corresponding grounding box. This advancement significantly enhances the efficiency of medical personnel.

Next, we first provide an overview of our architecture for medical phrase prediction and grounding box detection. We then review the multi-hypothesis prediction model and build upon it to establish an uncertainty-aware prediction model for grounding boxes. Furthermore, we present more details of the uncertainty-aware box decoder. Finally, we introduce the overall training loss function.

B. The Architecture of Medical Report Grounding

Medical phrase identification from medical report. Our primary objective is to generate medical phrases as input for predicting grounding boxes. However, being restricted solely to medical phrases may lead to unnecessarily complicating the framework as in [1]. The currently booming large-scale visual language models can predict text by taking images and text as output. Remarkably, building upon the insights from [39], we leverage a large visual language model to

generate phrases and utilize their embeddings as inputs for predicting grounding boxes, thereby enhancing the elegance of the framework. Consequently, given an X-ray image $\mathbf{x}_i$ and a medical report $\mathbf{x}_t$, we first obtain the objective function as follows:

$$\begin{cases} \mathbf{e} = \mathcal{F}_{vlm}\left(\mathbf{x}_i, \mathbf{x}_t\right) \\ \mathbf{y}_p = f_\theta(\mathbf{e}) \end{cases}. \tag{2}$$

Here, $\mathcal{F}_{\text{vlm}}$ denotes the large VLM equipped with LoRA [40] for parameter-efficient fine-tuning. LoRA adapters are inserted into the attention modules of the VLM backbone, specifically targeting the query and value projection layers with a rank $r = 8$ and a scaling factor $\alpha = 16$, while the original model weights are frozen and only the LoRA parameters are updated during training. The function $f_\theta(\cdot)$ represents the MLP projection layer, which predicts the phrase $\mathbf{y}_p$ based on the phrase prediction embedding $\mathbf{e}$. It is noteworthy that in Eq. 2, we employ the box embedding $\mathbf{e}_{box}$ in the $\mathbf{e}$ as a detection paradigm, integrating detection capabilities into the VLM. To facilitate this, we expand the vocabulary of the original language model by a new token, namely <BOX>, which serves as a marker for requesting detection box outputs. Next, we utilize the <BOX> token corresponding to the final-layer embedding of the VLM to predict the grounding box.

Phrase Grounding. First, we can directly use LLaVA equipped with LoRA to predict the initial grounding box $\mathbf{y}_b^0$, that is:

$$\mathbf{y}_b^0 = f_\Phi\left(\mathbf{e}_{box}\right), \tag{3}$$

This indicates that the <BOX> token is directly used as input to an MLP layer $f_\Phi(\cdot)$ for grounding box prediction. However, this grounding box may not effectively leverage the information from the box embedding $\mathbf{e}_{box}$. Therefore, to enhance the accuracy of detection grounding box, we adopt the widely-used SAM framework [41], utilizing the <BOX> embedding as a prompt for precise prediction. Specifically, we employ the visual backbone $\mathcal{F}_{enc}$ to extract visual features $\mathbf{z}_{enc}$. The box embedding $\mathbf{e_{box}}$ is then combined with the visual features and fed into the decoder $\mathcal{F}_{dec}$. Notably, $\mathbf{e_{box}}$ serves as a contextual cue. Compared with SAM framework [41], we enhance our model by modifying the final layer to box decoder, thereby obtaining grounding box embedding $\mathbf{z}_{dec}$. The process can be described as follows:

$$\begin{cases} \mathbf{z}_{enc} = \mathcal{F}_{enc}\left(\mathbf{x}_i\right) \\ \mathbf{z}_{dec} = \mathcal{F}_{dec}\left(\mathbf{z}_{enc}, \mathbf{e}_{box}\right) \end{cases}. \tag{4}$$

C. Multiple Hypothesis Prediction Model

To enhance the reliability of ground box predictions, we transition the prediction model from a single-point estimation framework to a multi-hypothesis estimation paradigm. In this context, we provide a brief review of multi-hypothesis prediction models. We assume the development of a prediction function capable of generating N predictions for a given input x:

$$\vec{f}_\Phi(x) = \left(f_\Phi^1(x), \ldots, f_\Phi^N(x)\right), \tag{5}$$

Where $\vec{f}_\Phi : \mathcal{X} \to \mathcal{Y}$ is a predictor, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ as the input and output variables. Then, calculating the loss $\mathcal{L}$ as always being the closest of the N predictions will minimize the following objective:

$$\int_{\mathcal{X}} \sum_{i=1}^{N} \int_{\mathcal{Y}_i(x)} (f_\Phi^i(x), y) p(x, y)\, dy\, dx, \tag{6}$$

$p(x, y) = p(y|x)p(x)$ denotes the joint probability density of the input variable x and the label y. We consider the Voronoi tessellation of the label space $\mathcal{Y} = \bigcup_{i=1}^{N} \mathcal{Y}_i$, which is induced by M generators $g_(x)$, and the loss function $\mathcal{L}$:

$$\mathcal{Y}_i(x) = \left\{ y \in \mathcal{Y} : (g^i(x), y) < (g^j(x), y)\ \forall j \neq i \right\}. \tag{7}$$

The Voronoi tessellation can be intuitively understood as partitioning the space such that each cell contains all points that are closest to a specific generator. In this case, the notion of closeness is determined by the loss function $\mathcal{L}$. Therefore, Eq. 6 partitions the space into N Voronoi cells, each corresponding to one of the predicted hypotheses $f_i(x)$, and aggregates the loss associated with each cell.

Theorem 1 For Eq. 6 to attain its minimum, a necessary condition is that the generators $g^j(x)$ coincide with the predictors $f_\theta^j(x)$, and both must correspond to a centroidal Voronoi tessellation:

$$g^i(x) = f_\Phi^i(x) = \frac{\int_{\mathcal{Y}_i} \mathcal{L}(f_\Phi^i(x), y) p(y|x)\, dy}{\int_{\mathcal{Y}_i} p(y|x)\, dy}. \tag{8}$$

For a more detailed proof and derivation of this theorem can be found in [42]. Intuitively, minimizing Eq. 6 aims to find the best piecewise constant approximation to the conditional distribution of labels in the output space. The hypotheses will divide the space into cells in such a way that the expected loss relative to their conditional means is minimized. Next, we will further explain the proposed uncertain prediction model by incorporating the multi-hypothesis prediction model.

D. Uncertain Prediction Model for Grounding Box

Directly predicting the grounding box using $\mathbf{z}_{\text{dec}}$ from Eq. 4 often results in a single deterministic output, which can suffer from overconfidence or insufficient reliability, making it difficult to gain acceptance from clinical experts in medical diagnostics. To address this limitation, we propose the **U**ncertain **P**rediction **M**odel (**UPM**). UPM incorporates the multi-hypothesis model described earlier, utilizing multiple hypothesis predictions to maintain output diversity, rather than relying solely on a single deterministic prediction model. More importantly, our model reveals the uncertainty (i.e., variance) of multiple hypothesis predictions, thereby overcoming unrealistic or blurred predictions in regression grounding box prediction. By incorporating this model into the vision output of LLMs, we provide a more reliable and robust grounding box prediction, which is crucial for gaining the trust and confidence of clinical experts in the field of medical diagnosis.

Specifically, instead of the expected value distribution for a single prediction $\mathbf{y}_b = f_\Phi\left(\mathbf{z}_{dec}\right)$, we aim to obtain multiple predictions to potentially achieve better approximations. To this end, we assume that the grounding box prediction is provided by N predictions:

$$\vec{f}_\Phi\left(\mathbf{z}_{dec}\right) = \left(f_\Phi^1\left(\mathbf{z}_{dec}\right), \ldots, f_\Phi^N\left(\mathbf{z}_{dec}\right)\right), \tag{9}$$

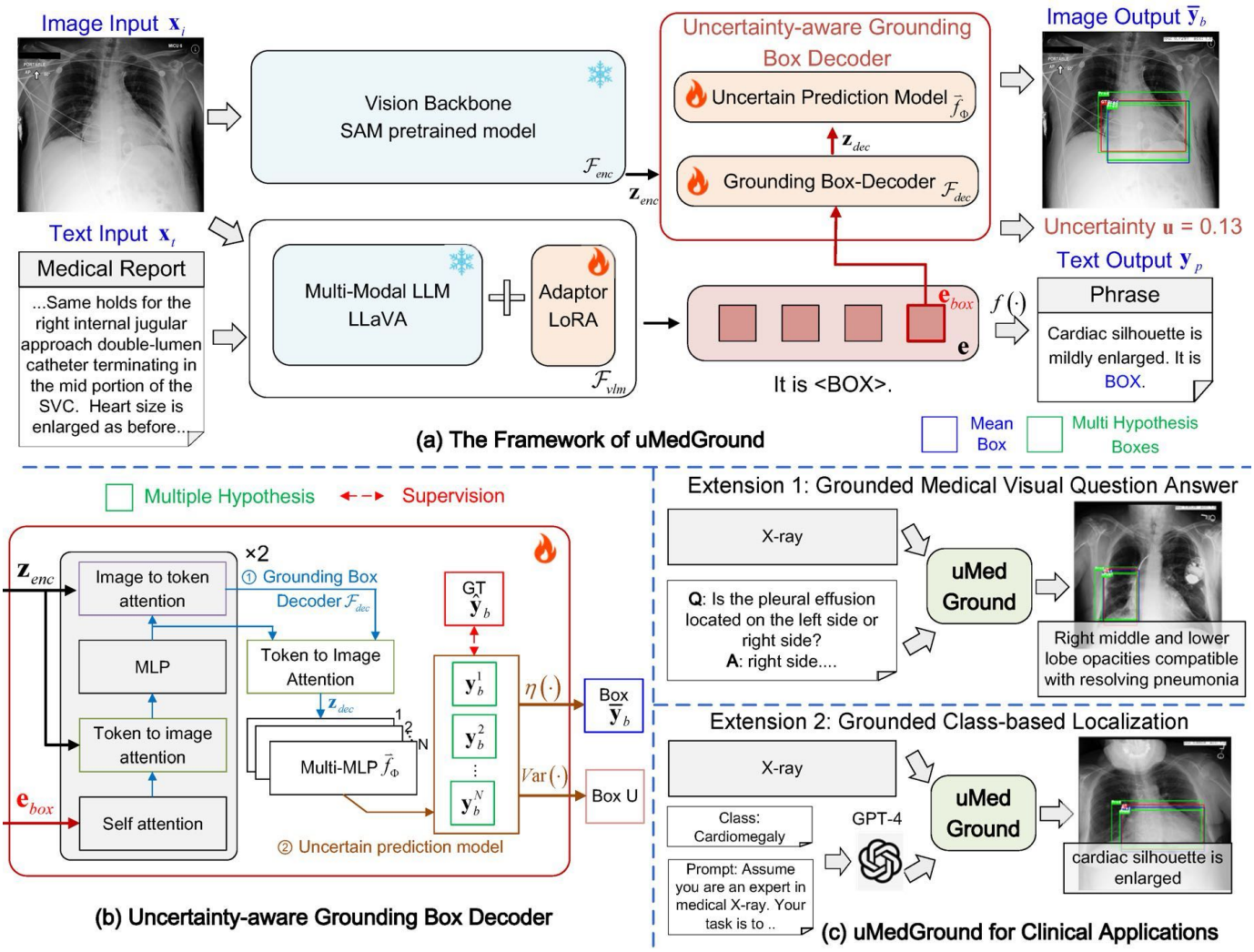

Fig. 2. Overview of **uMedGround**, an uncertainty-aware vision-language model designed for identifying and grounding medical diagnostic phrase. (a) The overall framework of uMedGround, which aligns medical reports with uncertainty-aware visual evidence at the phrase level. (b) The uncertainty-aware box decoder, which refines spatial grounding predictions based on multiple hypothesis predictions. (c) A potential clinical extension demonstrating how uMedGround can assist in explainable decision-making by linking grounded phrases with corresponding image regions.

where $f_\Phi^N(\mathbf{z}_{dec})$ represents the N-th prediction hypothesis derived from the embedding feature $\mathbf{z}_{dec}$. Next, similar to Eq. 6, we calculate that the uncertain-aware grounding box prediction loss $\mathcal{L}_u$ that is always the closest one out of N predictions, thus recommending minimization:

$$\sum_{i=1}^{N} \mathcal{L}_u\left(\hat{\mathbf{y}}_b, f_\Phi^i(\mathbf{z}_{dec})\right), \tag{10}$$

Then, we consider the Voronoi tessellation of the label space $\mathbf{y}_b = \cup_{i=1}^N \mathbf{y}_b^i$, which consists of multiple hypothesis predictions. Subsequently, we use the mean among the multiple hypothesis predictions as the final prediction and the variance as the uncertainty, which can be expressed as follows:

$$\bar{\mathbf{y}}_b = \eta\left(\sum_{i=1}^{N} \mathbf{y}_b^i\right), \mathbf{u} = Var\left(\sum_{i=1}^{N} \mathbf{y}_b^i\right), \tag{11}$$

where $\eta(\cdot)$ and $Var(\cdot)$ denote the mean and variance of their predictions.

E. Uncertainty-aware Grounding Box Decoder

To obtain uncertainty estimation for grounding box prediction, we introduce the uncertainty-aware grounding box decoder, detailed in Fig. 2 (b). Comprising two essential components, namely the grounding box decoder $\mathcal{F}_{dec}$ as described in Sec. III-B and the UPM $\sum_{i=1}^{N} f_\Phi^i$ detailed in Sec. III-D, this decoder primarily decodes the input image embedding and box embedding, producing a multi-modal feature embedding $\mathbf{z}_{dec}$. Additionally, UPM is employed for forecasting supervision and uncertainty estimation of joint embedding. Specifically, we employ multiple MLP layers $\vec{f}(\cdot) = (f_1(\cdot), f_2(\cdot), \ldots, f_N(\cdot))$, to obtain multiple hypothesis predictions $\sum_{i=1}^{N} \mathbf{y}_b^i$ of the grounding boxes. In the end, the parameters of the box distribution are employed in Eq. 11 to compute the predictions for both the grounding box and uncertainty. Therefore, the final expression of Eq. 1 can be

represented as:

$$(\mathbf{y}_p, \bar{\mathbf{y}}_b, \mathbf{u}) = \mathcal{F}(\mathbf{x}_t, \mathbf{x}_i). \tag{12}$$

F. Training to Form Medical Report Grounding

The proposed model is designed to be trained end-to-end and consists of three key components: phrase generation, detection box prediction, and uncertainty estimation.

Phrase identification. To enhance the ability of the visual language model to effectively summarize medical phrases in medical reports, we employ an auto-regressive cross-entropy loss. This loss function enables the model to learn the correlation between the predicted phrase and the actual phrase. Concretely, during the training process, given a real-valued phrase $\hat{\mathbf{y}}_p$, the formulation of the training process is as follows:

$$\mathcal{L}_p = \mathcal{L}_{CE}(\hat{\mathbf{y}}_p, \mathbf{y}_p). \tag{13}$$

Coarse phrase grounding. Furthermore, we supervise the initial prediction of the grounding box in Eq. 2 using two loss functions: smooth L1 loss [43] and generalized intersection over union (GIoU) loss [44]. These loss functions serve as guidance for improved box prediction. Specifically, during the optimization process, given a real box $\hat{\mathbf{y}}_b$, the formulation can be expressed as follows:

$$\mathcal{L}_b = \mathcal{L}_{l_1}(\hat{\mathbf{y}}_b, \mathbf{y}_b^0) + \mathcal{L}_{giou}(\hat{\mathbf{y}}_b, \mathbf{y}_b^0). \tag{14}$$

Uncertainty-aware phrase grounding. As indicated in Eq. 10, we utilize the uncertainty-aware prediction model to perform the final grounding box prediction. This optimization process combines the aforementioned L1 loss and GIoU loss, leading to the following equation:

$$\mathcal{L}_u = \sum_{i=1}^{N} \mathcal{L}_{l_1}(\hat{\mathbf{y}}_b, w_i \mathbf{y}_b^i) + \mathcal{L}_{giou}(\hat{\mathbf{y}}_b, w_i \mathbf{y}_b^i). \tag{15}$$

Unlike Eq. 10, we introduce a weight w_i to relax the hard minimization constraint in it. This idea stems from the fact highlighted in [45]: the predictor $f_\Phi(\cdot)$ might be initialized far from the target label $\hat{\mathbf{y}}_b$, such that all y fall into a single Voronoi cell. The weights are assigned following these principles:

$$w(\alpha) = \begin{cases} 1-\lambda, & \text{if } \alpha \text{ is true,} \\ \frac{\lambda}{N-1}, & \text{otherwise.} \end{cases} \tag{16}$$

The label y_b is assigned to the nearest hypothesis with a weight of $1-\lambda$, while the remaining hypotheses share a weight of $\frac{\lambda}{N-1}$. The sum of all weights $\sum_{i=1}^{N} w_i$ is constrained to 1.

Total objectives. Finally, the total target training function can be summarized as the summation of the aforementioned three components:

$$\mathcal{L}_{all} = \mathcal{L}_p + \mathcal{L}_b + \mathcal{L}_u \tag{17}$$

IV. Experiments

A. Experimental Settings

1) Curated Datasets: MRG-MS-CXR & MRG-ChestX-ray8 dataset. Quantitative assessments and benchmarks are crucial for grounding tasks in inferential medical reporting. To ensure reliable evaluation, we curated a benchmark dataset by collecting corresponding original reports from the MIMIC-CXR dataset [46]. It comprises 1153 Image-Phrase-BBox triplet samples. To ensure that each sample consists of an X-ray image, medical report, phrase query, and bounding box quadruple, we performed preprocessing on the MS-CXR dataset. We named it as MRG-MS-CXR data, which encompassed diverse X-ray report diagnostic scenarios, as depicted in Fig. 1 **b**. Its diagnostic phrases mainly include cardiomegaly, pneumothorax, pleural effusion, etc. The MRG-MS-CXR benchmark comprises a total of 835 pairs of reports, images, grounding boxes, and phrases. Similarly, we constructed pairs from the ChestX-ray8 dataset [47]. As the original ChestX-ray8 dataset only contains paired images, grounding boxes, and phrases, we used GPT-4 to expand each phrase into a corresponding report. This newly constructed paired dataset is referred to as MRG-ChestX-ray8. The MRG-ChestX-ray8 benchmark comprises a total of 984 paired reports, images, grounding boxes, and phrases. Following the [1], we split the two benchmarks into train-validation-test sets in a ratio of 7:1:2.

2) Clinical Application Datasets: MRG-MIMIC-VQA & MRG-MIMIC-Class dataset. To apply the proposed uMedGround to grounded medical visual question-and-answer (VQA) and class-based localization, we constructed sub-datasets based on the existing Medical-Diff-VQA [48] and MIMIC-CXR datasets. For medical VQA, we extracted all VQA instances corresponding to MRG-MS-CXR from the Medical-Diff-VQA dataset, treating them as text inputs for uMedGround. This subset of the test dataset is named MRG-MIMIC-VQA. Similarly, we extracted the categories most relevant to MRG-MS-CXR from the MIMIC-CXR dataset and used GPT-4's descriptions of these categories as text inputs for uMedGround. We excluded samples not present in both datasets, resulting in a total of 158 paired VQA samples and 163 class-based localization samples corresponding to the MRG-MS-CXR test set. This subset of the test dataset is referred to as MRG-MIMIC-Class.

3) Evaluation Metrics: To evaluate the effectiveness of the MRG task, we employ two types of evaluation metrics to measure the accuracy of phrase prediction and bounding box estimation. We adopt three widely used metrics from prior natural language generation (NLG) studies to evaluate the quality of predicted phrases. Specifically, BLEU measures n-gram precision between the generated and reference phrases; ROUGE-L captures recall-oriented similarity based on the longest common subsequence; and CIDEr emphasizes the relevance of informative phrases by incorporating term frequency-inverse document frequency weighting. To further evaluate the clinical and semantic fidelity of the predictions, we additionally incorporate three clinically meaningful metrics. CheXbert score [49] quantifies agreement with a CheXbert classifier across 14 clinical conditions, providing insight into diagnostic consistency. RadGraph F1 [50] assesses the overlap of predicted and ground-truth medical entities and relations within a structured radiology knowledge graph. Finally, BERTScore [51] leverages contextual embeddings to capture deeper semantic similarity beyond surface-level lexical matching. Additionally, we utilize the widely-used metric mIoU

(mean Intersection over Union) in bounding box prediction for a comprehensive comparison. Furthermore, we introduce the accuracy of bounding box prediction as an evaluation metric. Specifically, we consider a predicted bounding box as a positive sample if its mIoU with the ground truth bounding box exceeds certain thresholds (0.1, 0.3, and 0.5). We represent these thresholds as AP10 (mIOU >0.1), AP30 (mIOU >0.3), and AP50 (mIOU >0.5), respectively.

4) Baselines: We compare our proposed method with several general and medical visual grounding approaches, including RefTR [52], VGTR [3], TransVG [2], and MedRPG [1]. All methods are evaluated under the same settings as our model to ensure a fair comparison. Since these baselines require pre-extracted medical phrases as input, we use GPT-4 [53], LLaMA2 [54], and a fine-tuned version with LoRA adapter of LLaMA2-FT to extract phrases from medical reports during inference. In addition, we include comparisons with commonly used large language and vision-language models, such as GPT-4o [8], LLaVA-13B [9], and InternVL (specifically, InternVL2-8B) [10, 11]. Notably, both LLaVA-13B and InternVL were further fine-tuned using LoRA adapter on our task-specific dataset and are referred to as LLaVA-FT and InternVL-FT, respectively.For all baseline methods, we utilize their official implementations and ensure that training and evaluation are conducted under consistent experimental conditions.

5) Implementation Details: We conducted the following experiments on the PyTorch platform with an NVIDIA 48G A6000. The training scripts are based on deepspeed engine. We trained our model by the AdamW optimizer with the learning rate of 0.0003. Besides, the WarmupDecayLR strategy was employed for the learning rate scheduler and the its iterations are set to 100. The gradient accumulation step and batch size is set to 10 and 8, respectively. During training, the best checkpoints on the validation dataset of baselines and our model were reported with different random seeds.

B. Experimental Results on MRG-MS-CXR dataset

First, we conduct experiments on the MRG-MS-CXR dataset derived from MIMIC-CXR dataset [46], primarily demonstrating phrase identification, phrase grounding prediction, and visual comparison results.

1) Phrase identification: It is worth noting that existing visual grounding approaches are not well-suited for the end-to-end medical report grounding task, as they are not designed to identify diagnostic phrases, which presents a unique and critical challenge in the clinical domain. To address this issue, we explored a range of language models to extract diagnostic phrases directly from medical reports, including GPT-4 [8] and LLaMA2 [54] under a zero-shot setting. Additionally, we fine-tuned LLaMA2 using LoRA (referred to as LLaMA2-FT) and evaluated InternVL2-8B [10] on the same dataset to assess their capability in diagnostic phrase identification. Tab. I summarizes the diagnostic phrase extraction performance of GPT-4, LLaMA2, LLaMA2-FT, InternVL2-8B, and our proposed method. In the zero-shot setting, GPT-4 performs poorly in identifying key phrases, while LLaMA2 shows moderate improvement but still falls short of clinical requirements. Although fine-tuning with LoRA allows LLaMA2-FT to surpass InternVL2-8B in overall performance, both models remain constrained in their ability to accurately capture domain-specific diagnostic content. In contrast, our proposed approach, uMedGround, demonstrates significantly superior performance in diagnostic phrase recognition. It achieves substantial gains across standard NLP metrics (BLEU, ROUGE-L, and CIDEr) as well as three clinically relevant evaluation metrics (CheXbert score, RadGraph F1 and BERTScore). These improvements are primarily attributed to our tailored fine-tuning of LoRA layers within a multimodal LLM, which is specifically optimized for foundational medical report understanding tasks. By fine-tuning LLaVA, we equip the model with effective diagnostic phrase recognition capabilities and transform it into a fully end-to-end multimodal LLM framework. This design eliminates the need for the traditional two-stage grounding box prediction pipeline, simplifying and enhancing the grounding process.

2) Phrase Grounding: Table II presents the comparison of grounding box prediction performance on the MRG-MS-CXR dataset. Since most baseline methods do not support end-to-end box prediction, we first employ GPT-4 and LLaMA2 in a zero-shot setting to extract diagnostic phrases that may facilitate visual grounding. We compare the results obtained by applying MedRPG [1], TransVG [2], RefTR [52], and VGTR [3] using these phrases as input. Among them, TransVG demonstrates better performance than RefTR and VGTR, with MedRPG closely following. However, all of these methods achieve relatively low mIoU scores, which may be attributed to the limited quality and specificity of the phrases extracted from diagnostic reports. To further explore this, we apply a fine-tuned version of LLaMA2 with LoRA adapter to extract more accurate diagnostic phrases under the same dataset settings. As shown in the third major block of Tab. II, this significantly improves the performance of both general-purpose and medical visual grounding baselines. Notably, our proposed method, MedGround, consistently outperforms other state-of-the-art approaches. In particular, it achieves an AP30 above 70% and an AP10 close to 90%. Compared with TransVG using LLaMA2-FT for phrase input, our MedGround and uMedGround models achieve the highest prediction performance. The uMedGround variant further enhances MedGround by incorporating uncertainty-aware prediction, demonstrating its effectiveness in improving grounding accuracy in medical scenarios.

In addition, we compare our method with several recent VLMs, including GPT-4o [8], LLaVA-13B [9], and InternVL [10] (specifically InternVL2-8B). To ensure a fair comparison, both LLaVA-13B and InternVL2-8B were fine-tuned using the LoRA adapter under same dataset setting as our uMedGround model. These fine-tuned models are referred to as LLaVA-FT and InternVL-FT, respectively. As shown in Tab. II, both LLaVA-FT and InternVL-FT achieve notable improvements over the original GPT-4o, with InternVL-FT exhibiting the strongest overall performance among the baseline VLMs. Nevertheless, our proposed MedGround and uMedGround still outperform InternVL-FT by a significant margin. Specifically, MedGround and uMedGround surpass InternVL-FT by over 30% in AP50 and more than 20% in mIoU. The performance gains are mainly driven by the use

TABLE I
DIAGNOSTIC PHRASE IDENTIFICATION RESULTS ON MRG-MS-CXR DATASET WITH RESPECT TO BLEU_1 , BLEU_2, BLEU_3, BLEU_4, ROUGE_L, CIDER, CHEXBERT SCORE, RADGRAPH F1 AND BERTSCORE. OURS MEANS THE PROPOSED METHOD UMEDGROUND.

Method	BLEU_1	BLEU_2	BLEU_3	BLEU_4	ROUGE_L	CIDEr	CheXbert score	RadGraph F1	BERTScore
GPT-4	14.81	11.34	8.95	6.69	17.76	0.60	0.5560	0.1464	0.3336
LLaMA2	15.89	12.58	10.25	8.49	20.06	0.88	0.5619	0.1799	0.3793
LLaMA2-FT	36.68	31.95	28.75	26.60	41.48	2.65	0.7571	0.3145	0.5151
InternVL2-8B	33.34	28.11	24.49	21.79	20.06	2.06	0.7363	0.2558	0.4412
Ours	**82.79**	**82.11**	**79.04**	**77.80**	**65.90**	**5.16**	**0.9224**	**0.5590**	**0.6990**

of a multimodal large language model, which extracts richer diagnostic phrases and enables the fine-tuned visual decoder to jointly learn the $<$ BOX $>$ token and visual features, leading to more consistent and precise grounding.

TABLE II
GROUNDING BOX PREDICTION ON MRG-MS-CXR DATASET WITH RESPECT TO AP10 (ACC(MIOU$>$0.1)), AP30 (ACC(MIOU$>$0.3)), AND AP50 (ACC(MIOU$>$0.5)) AND MIOU. ZS AND FT MEAN THE ZERO SHOT AND FINE-TUNED, RESPECTIVELY. FT DENOTES THE FINE-TUNED MODEL.

	Method	AP10	AP30	AP50	mIOU
GPT-4	MedRPG [1]	57.49	40.12	18.56	25.73
	TransVG [2]	59.28	40.72	22.16	26.28
	RefTR [52]	56.29	32.34	16.17	23.22
	VGTR [3]	50.90	29.94	15.57	20.55
LLaMA2	MedRPG [1]	49.70	35.93	19.76	22.13
	TransVG [2]	52.10	34.13	16.77	22.31
	RefTR [52]	55.09	32.34	16.17	21.75
	VGTR [3]	47.31	28.74	14.37	19.11
LLaMA2-FT	MedRPG [1]	68.86	57.49	46.71	41.44
	TransVG [2]	71.86	58.68	46.71	42.60
	RefTR [52]	69.46	53.29	43.11	38.19
	VGTR [3]	72.46	51.50	41.92	38.06
End-to-end	GPT-4o [8]	24.55	8.98	2.99	7.51
	LLaVA-FT [9]	61.08	29.34	4.79	19.36
	InternVL-FT [10]	57.49	36.53	17.96	23.21
	MedGround	87.43	71.26	51.50	46.04
	uMedGround	**89.82**	**72.46**	**53.29**	**47.65**

3) Visual Comparisons: Figure 3 presents a visual comparison of three representative cases from the MRG-MS-CXR dataset, highlighting the grounding performance of our proposed method versus existing approaches. The focus is on evaluating the ability to localize key diagnostic phrases from medical reports. In Figures 3**b)** and **c)**, we show the results of representative existing methods guided by different LLMs. Figures 3**d)** and **e)** illustrate the performance of the fine-tuned InternVL model and our proposed uMedGround approach, respectively. In Case 1 (cardiomegaly), existing methods such as MedRPG [1] and TransVG [2] demonstrate poor grounding performance when guided by phrases extracted using GPT-4. When guided by phrases extracted via a LLaMA2 model fine-tuned with LoRA adapters, these methods show slightly improved localization performance, although the mIoU remains close to 10%. The fine-tuned InternVL model offers little improvement but still exhibits missed detections or false positives, achieving a much lower mIoU compared to our uMedGround, which reaches 83.28%. A similar trend is observed in Case 2. While MedRPG and TransVG again perform slightly better under the guidance of LLaMA2-FT compared to GPT-4, their performance remains suboptimal. The fine-tuned InternVL also underperforms, whereas uMedGround achieves an impressive mIoU of 74.96%. In Case 3, where the grounding task is relatively easier, all methods show improved localization. MedRPG and TransVG, guided by phrases from LoRA-tuned LLaMA2, outperform their GPT-4 counterparts. All four methods achieve approximately 70% mIoU, which is significantly better than the fine-tuned InternVL. Notably, MedRPG and TransVG generally outperform InternVL under LoRA-tuned LLaMA2 guidance due to their designs being more tightly coupled with visual-language grounding tasks. Nonetheless, our uMedGround approach still achieves the highest accuracy, with mIoU close to 75%. Overall, uMedGround not only delivers more accurate grounding box predictions but also reliably identifies diagnostic phrases. This effectiveness is largely attributed to the use of large language models fine-tuned with LoRA adapters for diagnostic phrase extraction, and their integration into the visual decoder for grounding. These findings demonstrate the feasibility and strong performance of our proposed medical report grounding framework, setting a new state-of-the-art and supporting the broader application of large language models in medical imaging tasks.

C. Experimental Results on MRG-ChestX-ray8 dataset

Furthermore, we conduct experiments on the MRG-ChestX-ray8 dataset derived from Chest-Xray8. Unlike the MRG-MS-CXR dataset, this dataset lacks original medical reports and primarily consists of simulated reports generated by GPT-4. Therefore, this section focuses on demonstrating phrase grounding and visual comparison results.

1) Phrase Grounding: Tab. III presents the comparison results of phrase grounding prediction on the MRG-ChestX-ray8 dataset. To adapt the ChestX-ray8 dataset to the medical report grounding task, we first utilized GPT-4 and LLaMA2 in a zero-shot setting to expand the original short phrases into medical-report-style inputs. We then evaluated several existing grounding methods, MedRPG [1], TransVG [2], RefTR [52], and VGTR [3], with the expanded medical reports as input. As shown in the first two major blocks of Tab. III, MedRPG

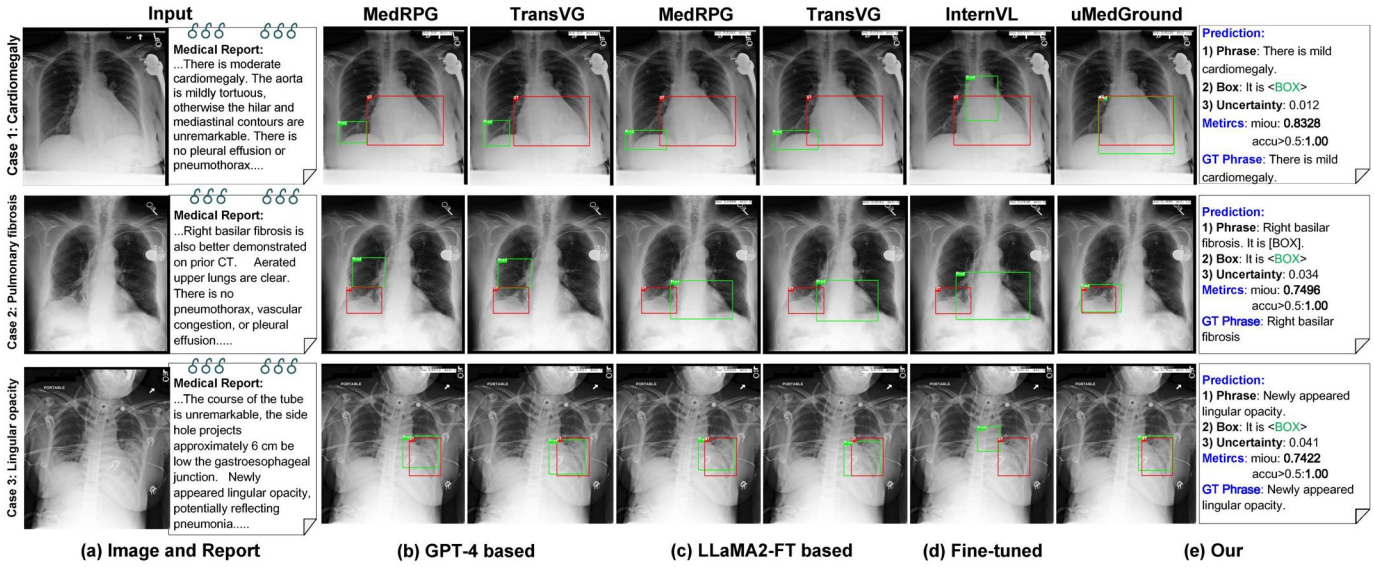

Fig. 3. Visual comparison of medical report grounding results across different methods on the MRG-ChestX-ray8 dataset. (a) Medical report and X-ray image input. (b) GPT-4 based methods for medical report grounding. (c) LLaMA2-FT based methods fine-tuned with LoRA for medical report grounding. (d) Fine-tuned method InterVL for medical report grounding. (e) Our proposed uMedGround method for medical report grounding.

consistently outperforms these methods under both GPT-4 and LLaMA2 expansions. To enable a fairer comparison, we further evaluated these methods using the original phrase annotations from MRG-ChestX-ray8 dataset, without any expansion into report-style inputs. The corresponding results are reported in the third major block of Table III, where we observe a notable improvement in phrase grounding accuracy. Despite these gains, our proposed models, MedGround and uMedGround, still outperform all existing baselines, particularly in AP metrics ranging from AP10 to AP50.

We also benchmarked end-to-end VLMs including GPT-4o, LLaVA-13B, and InternVL2-8B. All models received the same GPT-4 expanded medical reports as input to ensure consistency. GPT-4o was evaluated under a zero-shot setting, whereas LLaVA-13B and InternVL2-8B were fine-tuned using LoRA-based adapters under the same dataset setting as our method, and are thus denoted as LLaVA-FT and InternVL-FT, respectively. Among these VLMs, InternVL-FT achieved the best overall performance, reaching over 70% AP10, surpassing MedRPG under phrase-based input, but still underperformed in other metrics compared to MedRPG. Despite the improvements from recent VLMs, MedGround and uMedGround consistently achieved superior performance across all evaluation metrics. Notably, uMedGround outperformed the best-performing baseline by more than 10% in AP10 and 4% in mIoU. These improvements stem from our framework's capacity to extract key diagnostic phrases through a multimodal language model and integrate them into the visual decoder, which enhances grounding box consistency and relevance.

2) Visual Comparisons: Fig. 4 presents a visual comparison of representative methods on three cases from the MRG-ChestX-ray8 dataset. Fig. 4 **b)** and **c)** illustrate the visual grounding results of MedRPG and TransVG, using either GPT-4-augmented reports or ground-truth phrases as input.

TABLE III
GROUNDING BOX PREDICTION ON MRG-CHESTX-RAY8 DATASET WITH RESPECT TO AP10 (ACC(MIOU>0.1)), AP30 (ACC(MIOU>0.3)), AND AP50 (ACC(MIOU>0.5)) AND MIOU. FT DENOTES THE FINE-TUNED MODEL.

	Method	AP10	AP30	AP50	mIOU
GPT-4	MedRPG [1]	53.54	37.37	27.78	27.67
	TransVG [2]	54.04	37.37	25.76	26.27
	RefTR [52]	51.51	38.38	25.25	26.27
	VGTR [3]	56.06	27.78	9.60	19.28
LLaMA2	MedRPG [1]	50.51	34.85	24.24	25.53
	TransVG [2]	46.46	33.84	25.25	23.42
	RefTR [52]	58.08	33.33	13.64	21.38
	VGTR [3]	56.57	23.23	7.07	17.65
Phrase	MedRPG [1]	65.66	48.48	32.83	34.17
	TransVG [2]	63.13	45.96	31.31	32.48
	RefTR [52]	61.62	45.45	30.81	30.98
	VGTR [3]	63.13	46.97	29.29	29.9
End-to-end	GPT-4o [8]	25.76	18.69	6.06	10.12
	LLaVA-FT [9]	46.97	24.24	11.62	17.63
	InternVL-FT [10]	70.71	44.44	20.71	27.93
	MedGround	82.32	55.05	31.82	35.91
	uMedGround	**84.34**	**56.57**	**38.38**	**38.49**

Both methods generally exhibit suboptimal performance, often missing critical regions or producing false positives. Similar issues are observed in InternVL, a fine-tuned VLM on this task, as shown in the Fig. 4 **c)**. In the simpler Case 1 (cardiomegaly), TransVG, MedRPG, and InternVL methods approximately localize the correct region, with MedRPG demonstrating superior phrase grounding when the original phrase is provided as input. In contrast, uMedGround achieves

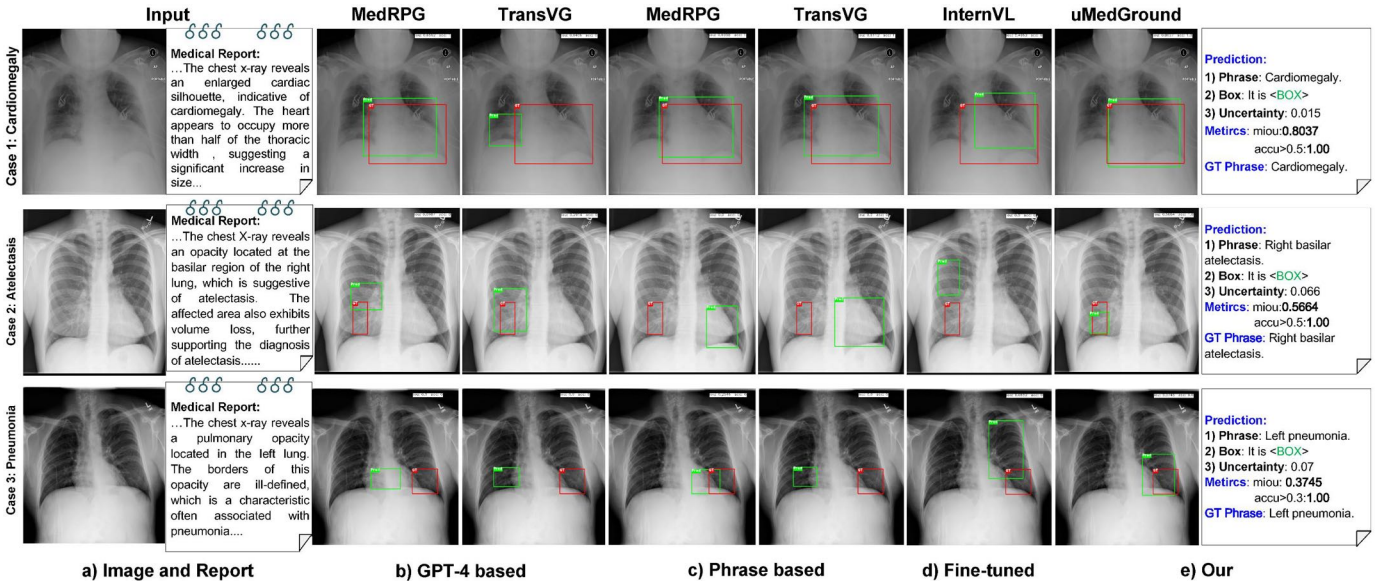

Fig. 4. Medical report grounding results of our method and competing baselines on the differennt cases. (a) Input medical report and corresponding chest X-ray image. (b) Grounding results using GPT-4-expanded medical reports. (c) Grounding based on ground-truth diagnostic phrases. (d) Grounding results from LoRA-fine-tuned InternVL (denoted as InternVL-FT). (e) Results of our proposed uMedGround method.

the closest alignment with the ground truth. In the more challenging Case 2 (Atelectasis), InternVL struggles to localize the target phrase, while MedRPG and TransVG with GPT-generated inputs succeed in identifying relevant regions but introduce false positive detections. When using ground-truth phrases as input, MedRPG and TransVG incorrectly focus on the cardiac area, likely due to data imbalance favoring that region. Despite missing certain regions, uMedGround identifies phrase-relevant areas that largely overlap with the annotated targets. Finally, in Case 3 (Pneumonia), TransVG performs poorly under both GPT-augmented and phrase-based inputs, while MedRPG shows modest improvement with phrase input but still underperforms compared to uMedGround. Overall, across diverse clinical scenarios, uMedGround consistently outperforms prior methods by more precisely aligning visual evidence with textual descriptions, resulting in improved visual grounding performance.

D. Performance of uMedGround on simulation Data

To evaluate the robustness of different grounding methods under low-quality image conditions, we introduced Gaussian noise into the original X-ray images. As shown in Table IV, we compared our proposed method with existing approaches under varying levels of Gaussian noise ($\sigma^2 = 0.01$ and $\sigma^2 = 0.1$). The results demonstrate that our uMedGround method exhibits superior robustness to noise. Specifically, on the MRG-MS-CXR dataset, under the guidance of GPT-4 extracted phrases, TransVG experienced a performance drop of 5.99% in AP30 and 3.76% in mIoU. MedRPG showed a similar level of degradation. When guided by phrases extracted from a LLaMA2 model fine-tuned with LoRA adapters, the performance drop of TransVG slightly improved to 4.19% (AP30) and 2.32% (mIoU), with MedRPG also exhibiting a moderate decline. In contrast, our uMedGround method showed only a 0.64% decrease in mIoU under the same noise conditions. On the MRG-ChestX-ray8 dataset, we applied a stronger level of Gaussian noise to further assess model stability. Both MedRPG and TransVG exhibited noticeable performance degradation under noisy conditions, regardless of whether the inputs were GPT-4 generated reports or manually extracted diagnostic phrases. In comparison, our uMedGround approach maintained stable performance, with only a 1.49% reduction in mIoU under high noise levels. Overall, the uncertainty estimation mechanism introduced in our medical report grounding framework, UPM, enables confidence prediction for the generated outputs and simultaneously enhances the model's robustness. This improvement can be attributed to UPM's multi-hypothesis design, which generates multiple plausible predictions and preserves output diversity, rather than relying solely on a single deterministic output.

In addition, we provide qualitative visualization results under different noise levels, as illustrated in Figure 5. Across both datasets, the introduction of Gaussian noise ($\sigma^2 = 0.01$ and $\sigma^2 = 0.1$) led to visibly degraded grounding box predictions in most baseline methods. Both TransVG and MedRPG were notably affected under various noise conditions. In contrast, our uMedGround model exhibited more stable and consistent grounding results, highlighting the advantage of incorporating uncertainty estimation.

E. Generalization Performance of uMedGround

To validate the generalization capability of the proposed model, we conducted a cross-dataset test. We used the model trained on the MRG-MS-CXR dataset to test the MRG-ChestX-ray8 dataset. Tab. V presents the results of different methods in terms of generalization performance. Although

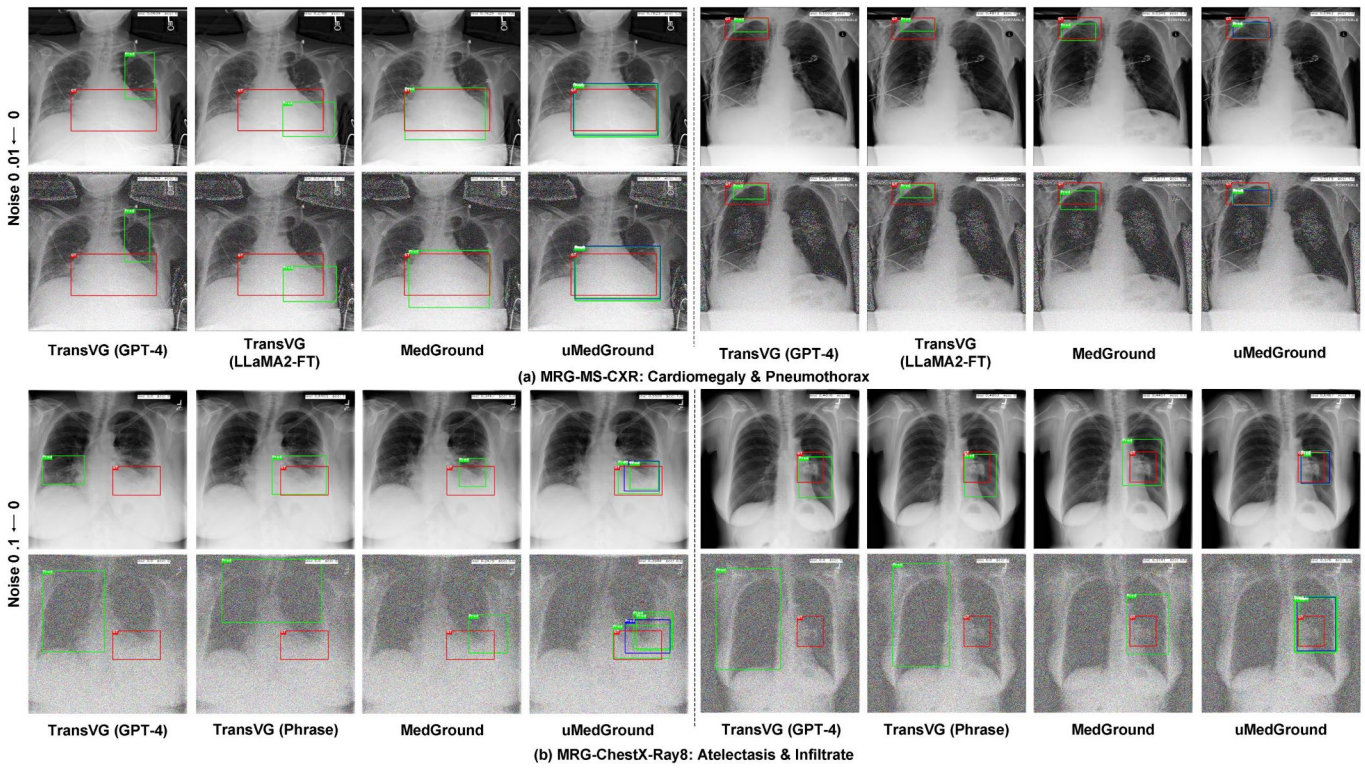

Fig. 5. (a–b) Grounding results under normal and Gaussian noise conditions on the MRG-MS-CXR and MRG-ChestX-ray8 datasets using TransVG (GPT-4), TransVG (LLaMA2-FT), TransVG (Phrase), MedGround, and uMedGround. In (a), TransVG (GPT-4) and TransVG (LLaMA2-FT) use diagnostic phrases extracted from medical reports by GPT-4 and LoRA-fine-tuned LLaMA2, respectively. In (b), TransVG (GPT-4) is based on GPT-4-expanded medical reports, while TransVG (Phrase) uses ground-truth diagnostic phrases extracted from the reports.

TABLE IV
GROUNDING BOX PREDICTION RESULTS OF DIFFERENT METHODS ON LOW-QUALITY IMAGES: MRG-MS-CXR (GAUSSIAN NOISE $\sigma^2 = 0.01$) AND MRG-CHESTX-RAY8 (GAUSSIAN NOISE $\sigma^2 = 0.1$).

	Method	MRG-MS-CXR		MRG-ChestX-ray8	
		AP30	mIOU	AP30	mIOU
GPT-4	MedRPG [1]	27.54	19.92	9.09	7.84
	TransVG [2]	34.73	22.52	25.76	16.69
FT/P	MedRPG [1]	53.89	39.82	10.61	9.69
	TransVG [2]	54.49	40.28	25.76	18.74
Ours	MedGround	70.06	44.65	48.48	32.52
	uMedGround	**72.46**	**47.01**	**55.05**	**35.16**

our proposed method shows lower performance in the AP50 metric compared to existing methods such as MedRPG and TransVG, it outperforms them in other metrics, including mIOU. Notably, the AP10 metric is 9.09% higher than the best existing method. This demonstrates that our model can promptly predict grounding boxes on other datasets with lower IOU thresholds, which is crucial for alerting clinical staff to abnormal conditions.

TABLE V
GENERALIZATION OF GROUNDING BOX PREDICTION ON MRG-CHESTX-RAY8 DATASET (TRAINING ON MRG-MS-CXR DATASET) WITH RESPECT TO AP10 (ACC(MIOU>0.1)), AP30 (ACC(MIOU>0.3)), AND AP50 (ACC(MIOU>0.5)) AND MIOU.

Method	AP10	AP30	AP50	mIOU
MedRPG [1]	66.16	40.40	26.26	28.35
TransVG [2]	62.63	39.39	23.74	26.63
MedGround	73.23	40.91	21.72	28.38
uMedGround	**75.25**	**42.42**	**23.74**	**30.43**

F. Clinical Extension of uMedGround

As shown in Fig. 2 **c**, we aim to leverage the extension of uMedGround for broader potential clinical applications. To this end, we applied it to primary healthcare VQA tasks and class-based localization tasks. For medical VQA, we used QA pairs as the textual input to uMedGround, demonstrating that medical QA can effectively serve as input for our model. Furthermore, in more complex clinical scenarios, physicians may provide only a textual category rather than detailed diagnostic insights. Based on this observation, we used category-based text as the textual input to uMedGround to show that it can also function as effective input for our model. Consequently, we constructed two

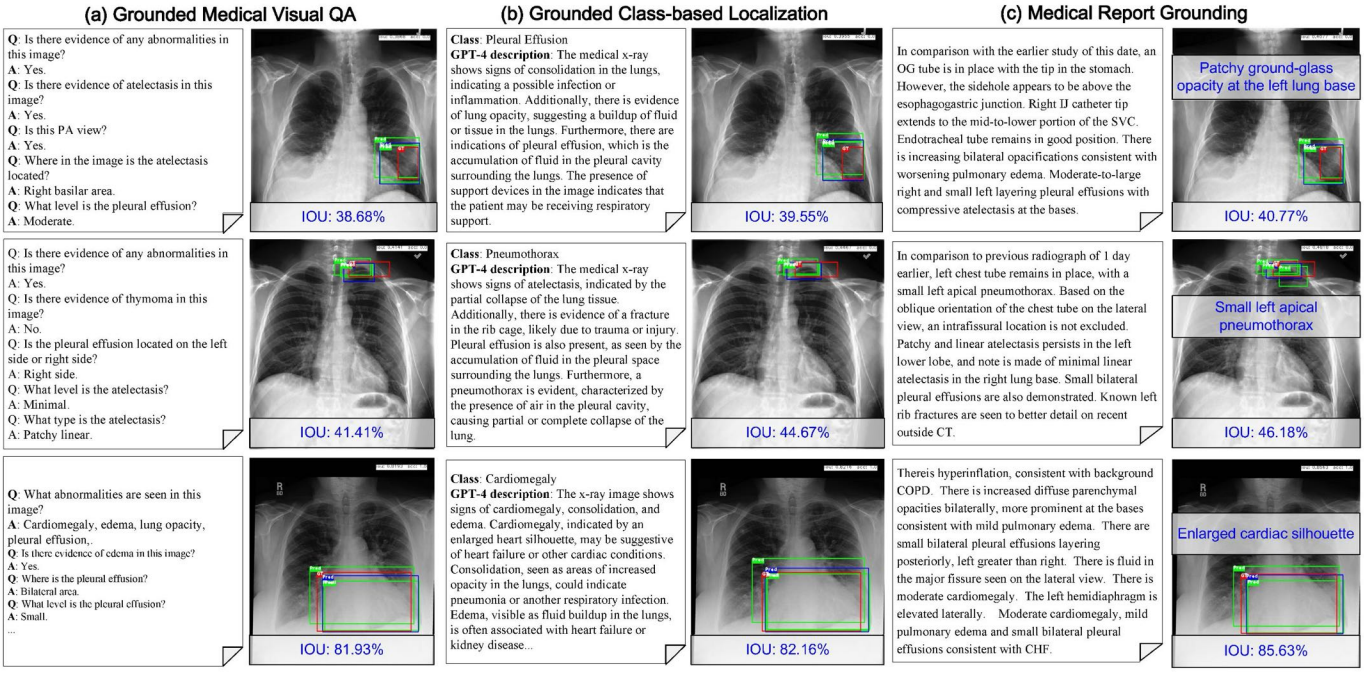

Fig. 6. Visual results on clinical applications. (a) Grounded medical visual question-and-answer (Medical question-and-answer as as text input). (b) Grounded Class-based localization (Class-based words with GPT-4 description as as text input). (c) Medical Report Grounding (Original report as text input).

clinical extension application sub-test datasets, MRG-MIMIC-VQA and MRG-MIMIC-Class, for experimentation.

The experimental results are summarized in Table VI. By observing the first and second rows, it is evident that the uMedGround model performs significantly better with VQA-based report descriptions generated by GPT-4 compared to class-based inputs. This improvement may be due to the fact that VQA-based descriptions are more specific than class-based inputs and may include location information. Furthermore, when compared to using the original report as input, the results show that the original report is better than the VQA-based description.

In addition, Figure 6 presents the visualization results of grounding box predictions under different cases with varying textual inputs. As shown in Figure 6(a), uMedGround successfully predicts the phrases and their corresponding grounding boxes. Furthermore, when class information is integrated as the textual input to uMedGround, the model's predictions for phrases and grounding boxes are consistent and outperform those using QA-based inputs. Finally, when the original report is used as the textual input to uMedGround, the model achieves the best results for both phrase and grounding box predictions. The above experiments validate the extension of uMedGround in clinical applications, demonstrating that even when QA-based inputs are used, the proposed model may still achieve pre-satisfactory results.

G. Ablation Study and Analysis

In this section, we present an ablation study to analyze the impact of different components within the proposed uMedGround framework. As shown in Tab. VII, uMedGround

TABLE VI
QUESTION-AND-ANSWER PAIRS, CLASS DESCRIPTIONS GENERATED BY GPT-4, AND ORIGINAL MEDICAL REPORT AS QUANTITATIVE RESULTS FOR UMEDGROUND.

Text Input	AP10	AP30	AP50	mIOU
MRG-MIMIC-Class	87.97	**73.42**	49.37	46.61
MRG-MIMIC-VQA	88.47	73.39	50.55	47.39
MRG-MS-CXR	**89.82**	72.46	**53.29**	**47.65**

consists of three core modules: LLaVA, SAM, and the uncertain prediction model.

A natural approach to predict grounding boxes is to leverage LLaVA by embedding a special token $< BOX >$ and directly regressing its coordinates. Due to GPU memory limitations, we adopt LoRA-based fine-tuning for LLaVA. As shown in the first row of Table VII, this configuration yields limited performance in box prediction. This may be attributed to the fact that LoRA tuning mainly enhances the language component for medical phrase extraction, while contributing less to visual feature learning. To improve visual representation, we introduce the SAM [55] as the visual backbone. We investigate three training strategies: training only the decoder while freezing the encoder, applying LoRA to the encoder, and training both the encoder and decoder. The result from training only the decoder is presented in the second row of Table VII, which clearly outperforms the LLaVA-only setup. This demonstrates the strong capability of SAM, based on the ViT-H architecture, in extracting visual cues relevant to medical grounding. We further examine how these three SAM configurations interact with the uncertainty-aware prediction module, which produces multiple hypotheses to improve prediction reliability. The

results are shown in the third to fifth rows of Table VII. Among these, the best performance is observed when the SAM encoder is kept frozen and only the decoder is trained. This suggests that fine-tuning the encoder in the medical domain may lead to overfitting, especially when both encoder and decoder are updated simultaneously. At last, the best-performing configuration, which we refer to as uMedGround, combines LoRA-tuned LLaVA, SAM with a frozen encoder and a trained decoder, and the uncertainty-aware prediction module. This combination effectively integrates multimodal language understanding with robust visual grounding, leading to improved accuracy and robustness in medical report grounding tasks.

TABLE VII
ABLATION STUDIES EXAMINING THE IMPACT OF DIFFERENT MODULES ON GROUNDING BOX PREDICTIONS. ✓ INDICATES THE MODULE IS INCLUDED. NUMBERS IN **BOLD FONT** INDICATE THE BEST RESULTS. D_{FT} DENOTES FINE-TUNING ONLY THE DECODER; E_{lr}+D_{FT} DENOTES FINE-TUNING THE ENCODER WITH THE LORA ADAPTER TOGETHER WITH FULL DECODER FINE-TUNING; E_{FT}+D_{FT} DENOTES FULL FINE-TUNING OF BOTH THE ENCODER AND DECODER.

LLaVA+LoRA	SAM			UPM	AP30	mIOU
	D_{FT}	E_{lr}+D_{FT}	E_{FT}+D_{FT}			
✓					29.34	19.36
✓	✓				71.26	46.04
✓	✓			✓	**72.46**	**47.65**
✓		✓		✓	67.07	45.31
✓			✓	✓	66.10	41.53

V. DISCUSSION

The proposed uMedGround framework demonstrates consistent and significant improvements across multiple datasets and evaluation dimensions. On the MRG-MS-CXR benchmark, our model outperforms various baselines not only on standard NLG metrics such as BLEU, ROUGE-L, and CIDEr, but also on clinically meaningful metrics including CheXbert score, RadGraph F1, and BERTScore. For the phrase grounding task, uMedGround surpasses both general visual grounding and specialized medical phrase grounding approaches. Even when using diagnostic phrases extracted by fine-tuned LLaMA, our model consistently achieves better performance. Moreover, uMedGround outperforms leading end-to-end vision-language models such as GPT-4o, LLaVA-13B, and fine-tuned InternVL, highlighting its ability to generate semantically accurate and clinically aligned diagnostic phrases and grounding outputs.

Similar findings hold on the MRG-ChestX-ray8 dataset, where uMedGround continues to deliver superior performance, especially when diagnostic phrases are used as input to other non end-to-end grounding methods. This consistency across datasets underscores the framework's robust capabilities in both differential diagnostic phrase identification and visual grounding.

Qualitative visualizations further illustrate that uMedGround produces more precise and semantically meaningful bounding boxes than existing medical phrase grounding methods. The multi-hypothesis grounding module improves reliability by modeling local uncertainty, which is crucial for building clinician trust. This feature is especially valuable in complex or low-quality chest X-rays, where deterministic models may produce overconfident but incorrect predictions. Experiments on synthetic data confirm that uMedGround maintains robustness under image degradation, ensuring consistent grounding performance. Additionally, cross-validation across two benchmarks demonstrates strong transferability with minimal performance degradation, further supporting its potential for clinical deployment.

We also extended uMedGround to real-world clinical tasks such as primary care VQA and class-based localization. These experiments show that our model can flexibly handle various forms of textual input, including free-text reports, VQA queries, and class labels, and accurately identify the corresponding visual evidence (grounding box). This versatility is critical for supporting clinicians in interpreting diverse types of input within practical workflows.

Despite these encouraging results, several limitations remain. First, although the multi-hypothesis grounding module is designed to be lightweight, it introduces a slight computational overhead during inference, which may pose challenges in resource-constrained settings. Second, performance may decline when the model encounters rare or novel diagnostic phrases that lack well-defined spatial representations in the training data. Future work could address this by integrating external retrieval-augmented modules to expand semantic coverage. Moreover, while our current evaluation relies on automated metrics, it is necessary to validate the framework's utility through prospective clinical studies, such as expert-in-the-loop evaluations in real-world settings.

In summary, uMedGround achieves a balanced integration of model performance, trustworthiness, and clinical utility. Its architecture, which features uncertainty-aware prediction and minimal reliance on handcrafted phrase annotations, makes it particularly well suited for deployment in real-world radiology workflows. The framework supports a range of tasks including diagnostic report generation, clinical question answering, and class-based phrase localization, demonstrating its versatility and potential to serve as a core component of multimodal medical AI systems in practical clinical settings.

VI. CONCLUSIONS

In conclusion, we introduce the MRG task as a novel approach to enhance medical image analysis and radiological diagnosis. Our proposed uMedGround model combines a multimodal LLM with a specially designed <BOX> token for embedding-based detection and integrates uncertainty-aware prediction to boost grounding box confidence and reliability. Extensive experiments demonstrate that uMedGround consistently outperforms state-of-the-art visual grounding methods and vision-language models in terms of both accuracy and robustness. Furthermore, we investigate the clinical applicability of various text inputs for grounding, showcasing the adaptability of our approach to diverse medical scenarios.

This work represents a significant advancement in multimodal medical image analysis, addressing critical challenges in radiological diagnosis by providing precise and interpretable grounding results. By bridging the gap between textual medical

reports and visual cues in medical images, our approach lays the groundwork for improved diagnostic workflows. In future work, we plan to extend our research in two directions. First, we will construct larger and more generalizable datasets to further evaluate the robustness of our method [56]. Second, we aim to integrate the system into real-world clinical workflows by collaborating with resident physicians to conduct human evaluations, exploring its potential for deployment in routine practice [57].

REFERENCES

[1] Z. Chen, Y. Zhou, A. Tran, J. Zhao, and et al., "Medical phrase grounding with region-phrase context contrastive alignment," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Cham: Springer Nature Switzerland, 2023, pp. 371–381.

[2] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, "Transvg: End-to-end visual grounding with transformers," in *ICCV*, 2021, pp. 1769–1779.

[3] Y. Du, Z. Fu, Q. Liu, and Y. Wang, "Visual grounding with transformers," in *ICME*. IEEE, 2022, pp. 1–6.

[4] C. Zhu, Y. Zhou, Y. Shen, G. Luo, X. Pan, M. Lin, C. Chen, L. Cao, X. Sun, and R. Ji, "Seqtr: A simple yet universal network for visual grounding," in *ECCV*. Springer, 2022, pp. 598–615.

[5] L. Zhou, Z. Zhou, K. Mao, and Z. He, "Joint visual grounding and tracking with natural language specification," in *CVPR*, 2023, pp. 23 151–23 160.

[6] R. He, P. Cascante-Bonilla, Z. Yang, A. C. Berg, and V. Ordonez, "Improved visual grounding through self-consistent explanations," in *CVPR*, 2024, pp. 13 095–13 105.

[7] A. Ichinose, T. Hatsutani, K. Nakamura, Y. Kitamura, S. Iizuka, E. Simo-Serra, S. Kido, and N. Tomiyama, "Visual grounding of whole radiology reports for 3d ct images," in *MICCAI*. Springer, 2023, pp. 611–621.

[8] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *NeurIPS*, vol. 36, 2024.

[10] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *CVPR*, 2024, pp. 24 185–24 198.

[11] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma *et al.*, "How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites," *arXiv preprint arXiv:2404.16821*, 2024.

[12] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Medical image analysis*, vol. 59, p. 101557, 2020.

[13] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *MICCAI*. Springer, 2019, pp. 605–613.

[14] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, "Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 61–72.

[15] L. Huang, S. Ruan, P. Decazes, and T. Denoeux, "Lymphoma segmentation from 3d pet-ct images using a deep evidential network," *International Journal of Approximate Reasoning*, vol. 149, pp. 39–60, 2022.

[16] K. Zou, X. Yuan, X. Shen, M. Wang, and H. Fu, "Tbrats: Trusted brain tumor segmentation," in *MICCAI*. Springer, 2022, pp. 503–513.

[17] M. Wang, T. Lin, L. Wang, A. Lin, K. Zou, X. Xu, Y. Zhou, Y. Peng, Q. Meng, Y. Qian *et al.*, "Uncertainty-inspired open set learning for retinal anomaly identification," *Nature Communications*, vol. 14, no. 1, p. 6757, 2023.

[18] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *NeurIPS*, 2018, pp. 3183–3193.

[19] J. Mukhoti, J. van Amersfoort, P. H. Torr, and Y. Gal, "Deep deterministic uncertainty for semantic segmentation," in *International Conference on Machine Learning Workshop on Uncertainty and Robustness in Deep Learning*, 2021.

[20] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *ICML*. PMLR, 2020, pp. 9690–9700.

[21] B. A. Plummer, P. Kordas, M. H. Kiapour, S. Zheng, R. Piramuthu, and S. Lazebnik, "Conditional image-text embedding networks," in *ECCV*, 2018, pp. 249–264.

[22] J. Wang and L. Specia, "Phrase localization without paired training examples," in *ICCV*, 2019, pp. 4663–4672.

[23] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi, "Multi-modal understanding and generation for medical images and text via vision-language pre-training," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6070–6080, 2022.

[24] F. Wang, Y. Zhou, S. Wang, V. Vardhanabhuti, and L. Yu, "Multi-granularity cross-modal alignment for generalized medical visual representation learning," *NeurIPS*, vol. 35, pp. 33 536–33 549, 2022.

[25] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Medklip: Medical knowledge enhanced language-image pre-training," *ICCV*, 2023.

[26] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, and O. Oktay, "Making the most of text semantics to improve biomedical vision–language processing," in *ECCV*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 1–21.

[27] K. Vilouras, P. Sanchez, A. Q. O'Neil, and S. A. Tsaftaris, "Zero-shot medical phrase grounding with off-the-shelf diffusion models," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[28] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen, "Chatcad: Interactive computer-aided diagnosis on medical image using large language models," *arXiv preprint arXiv:2302.07257*, 2023.

[29] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar, "Med-flamingo: a multimodal medical few-shot learner," in *Machine Learning for Health (ML4H)*. PMLR, 2023, pp. 353–367.

[30] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Towards generalist foundation model for radiology," *arXiv preprint arXiv:2308.02463*, 2023.

[31] W. Zhao, Z. Deng, S. Yadav, and P. S. Yu, "Heterogeneous knowledge grounding for medical question answering with retrieval augmented large language model," in *Companion Proceedings of the ACM Web Conference 2024*, 2024, pp. 1590–1594.

[32] Y. Gao, R. Li, E. Croxford, S. Tesch, D. To, J. Caskey, B. W. Patterson, M. M. Churpek, T. Miller, D. Dligach *et al.*, "Large language models and medical knowledge grounding for diagnosis prediction," *medRxiv*, pp. 2023–11, 2023.

[33] S. Bannur, K. Bouzid, D. C. Castro, A. Schwaighofer, S. Bond-Taylor, M. Ilse, F. Pérez-García, V. Salvatelli, H. Sharma, F. Meissen *et al.*, "Maira-2: Grounded radiology report generation," *arXiv preprint arXiv:2406.04449*, 2024.

[34] L. Huang, S. Ruan, Y. Xing, and M. Feng, "A review of uncertainty quantification in medical image analysis: probabilistic and non-probabilistic methods," *Medical Image Analysis*, p. 103223, 2024.

[35] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, and O. Ronneberger, "A probabilistic u-net for segmentation of ambiguous images," *NeurIPS*, vol. 31, 2018.

[36] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 3868–3878, 2020.

[37] Y. Luo, Q. Yang, Y. Fan, H. Qi, and M. Xia, "Measurement guidance in diffusion models: Insight from medical image synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[38] L. Huang, S. Ruan, P. Decazes, and T. Denœux, "Deep evidential fusion with uncertainty quantification and reliability learning for multimodal medical image segmentation," *Information Fusion*, vol. 113, p. 102648, 2025.

[39] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," *arXiv preprint arXiv:2308.00692*, 2023.

[40] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[41] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[42] Q. Du, V. Faber, and M. Gunzburger, "Centroidal voronoi tessellations: Applications and algorithms," *SIAM review*, vol. 41, no. 4, pp. 637–676, 1999.

[43] R. Girshick, "Fast r-cnn," in *ICCV*, 2015, pp. 1440–1448.

[44] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *CVPR*, 2019, pp. 658–666.

[45] C. Rupprecht, I. Laina, R. DiPietro, M. Baust, F. Tombari, N. Navab, and G. D. Hager, "Learning in an uncertain world: Representing ambiguity through multiple hypotheses," in *ICCV*, 2017, pp. 3591–3600.

[46] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, 2019.

[47] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *CVPR*, 2017, pp. 2097–2106.

[48] X. Hu, L. Gu, Q. An, M. Zhang, L. Liu, K. Kobayashi, T. Harada, R. M. Summers, and Y. Zhu, "Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering," in *ACM KDD*, 2023, pp. 4156–4165.

[49] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[50] S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng *et al.*, "Radgraph: Extracting clinical entities and relations from radiology reports," *arXiv preprint arXiv:2106.14463*, 2021.

[51] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert," *arXiv preprint arXiv:2004.09167*, 2020.

[52] L. Muchen and S. Leonid, "Referring transformer: A one-step approach to multi-task visual grounding," in *NeurIPS*, 2021.

[53] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," *arXiv preprint arXiv:2303.13375*, 2023.

[54] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[55] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[56] B. Liu, K. Zou, L. Zhan, Z. Lu, X. Dong, Y. Chen, C. Xie,

J. Cao, X.-M. Wu, and H. Fu, "Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis," *arXiv preprint arXiv:2411.16778*, 2024.

[57] M. Wang, T. Lin, A. Lin, K. Yu, Y. Peng, L. Wang, C. Chen, K. Zou, H. Liang, M. Chen *et al.*, "Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model," *Nature Communications*, vol. 16, no. 1, p. 5528, 2025.