Describe Anything in Medical Images

Xi Xiao [† 1] **Yunbei Zhang** [† 2] **Thanh-Huy Nguyen** [† 3] **Ba-Thinh Lam** [4] **Janet Wang** [2] **Lin Zhao** [5]
Jihun Hamm [2] **Tianyang Wang** [1] **Xingjian Li** [3] **Xiao Wang** [6] **Hao Xu** [7] **Tianming Liu** [* 8] **Min Xu** [* 3]

Abstract

Localized image captioning has made significant progress with models like the Describe Anything Model (DAM), which can generate detailed region-specific descriptions without explicit region-text supervision. However, such capabilities have yet to be widely applied to specialized domains like medical imaging, where diagnostic interpretation relies on subtle regional findings rather than global understanding. To mitigate this gap, we propose **MedDAM**, the first comprehensive framework leveraging large vision-language models for *region-specific captioning in medical images*. MedDAM employs medical expert-designed prompts tailored to specific imaging modalities and establishes a robust evaluation benchmark comprising a customized assessment protocol, data pre-processing pipeline, and specialized QA template library. This benchmark evaluates both MedDAM and other adaptable large vision-language models, focusing on clinical factuality through attribute-level verification tasks—elegantly circumventing the absence of ground-truth region-caption pairs in medical datasets. Extensive experiments on the VinDr-CXR, LIDC-IDRI, and SkinCon datasets demonstrate MedDAM's superiority over leading peers (including GPT-4o, Claude 3.7 Sonnet, LLaMA-3.2 Vision, Qwen2.5-VL, GPT-4RoI, and OMG-LLaVA) in the task, revealing the importance of region-level semantic alignment in medical image understanding and establishing MedDAM as a promising foundation for clinical vision-language integration.

[†]Equal contribution [1]University of Alabama at Birmingham [2]Tulane University [3]Carnegie Mellon University [4]AI VIETNAM [5]Northeastern University [6]Oak Ridge National Laboratory [7]Harvard Medical School [8]University of Georgia. Correspondence to: Tianming Liu <tliu@cs.uga.edu>, Min Xu <mxu1@cs.cmu.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

Vision-language models (VLMs) have made remarkable strides in generating natural language descriptions of visual content. Traditional captioning models often focus on interpreting entire images (Vinyals et al., 2015; Anderson et al., 2018), but many real-world applications require a deep understanding and precise descriptions of fine-grained and localized regions. Recent advances in region-level captioning (Johnson et al., 2016; Li et al., 2022; 2023; Lian et al., 2025) have introduced models capable of producing detailed region-specific descriptions without relying on explicit region-text supervision. Among them, the Describe Anything Model (DAM) (Lian et al., 2025) is particularly notable, leveraging focal prompting, a localized visual backbone, and a self-supervised data pipeline (DLC-SDP) to achieve state-of-the-art performance on localized captioning tasks for natural images.

Though promising for natural images, the extension of such models to medical images remains unexplored. In clinical practice, diagnostic interpretation could highly rely on subtle localized findings rather than a holistic understanding. Detailed descriptions of localized regions—capturing crucial features such as location, morphology, density, and boundary characteristics—are essential for accurate diagnosis and effective communication among healthcare professionals. However, existing medical image captioning methods (Jing et al., 2018; Chen et al., 2020; Wang et al., 2022; Huang et al., 2023) are developed to generate image-level descriptions, failing to deliver fine-grained region-specific descriptions urgently needed by clinicians (e.g., radiologists) for diagnosis.

Unlike segmentation models such as MedSAM (Ma et al., 2023) and MedSAM2 (Ma et al., 2024) that predict explicit region masks, the recent breakthrough DAM (Lian et al., 2025) generates free-form textual descriptions without predefined classes or structured labels (Wang et al., 2024b; Xiao et al., 2024). However, applying DAM directly to medical images introduces unique challenges. Moreover, medical datasets rarely provide region-specific captions, rendering conventional captioning evaluation metrics (e.g., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004)) inapplicable. This raises a critical question: *Can a general localized caption-*

ing model like DAM generalize to medical images, and how should its outputs be evaluated in the absence of ground-truth region-specific descriptions?

To answer it, we introduce our **MedDAM** framework, a practical solution for region-specific captioning across diverse medical images. It integrates three key components: (1) medical expert-designed prompts tailored for different organs and imaging modalities (chest X-ray, lung CT scan, dermatology, etc.), enabling appropriate domain-specific query; (2) a flexible region-of-interest detection pipeline that leverages existing segmentation models when boundary boxes or masks are unavailable in the original dataset, making it adaptable to any medical imaging collection; and (3) a specialized evaluation benchmark that assesses the quality and accuracy of detailed localized captioning without requiring reference captions through attribute-level verification tasks. This unified framework performs effectively across any available medical imaging datasets regardless of domain or availability of pre-existing segmentation annotations or captions, offering a versatile and robust framework for advancing region-specific understanding in medical images. We conduct extensive evaluations on three clinically significant datasets: VinDr-CXR for chest radiography, LIDC-IDRI for lung CT imaging, and SkinCon for skin imaging, covering critical diagnostic tasks across different imaging modalities. Our main contributions are summarized as follows:

- We introduce **MedDAM**, a framework that enables region-specific captioning for medical images through expert-designed prompts and a specialized evaluation protocol for clinical applications.
- We establish the first benchmark comparing MedDAM against leading large VLMs across chest radiography, lung CT, and dermatology, demonstrating its significant advantages in region-specific clinical understanding.

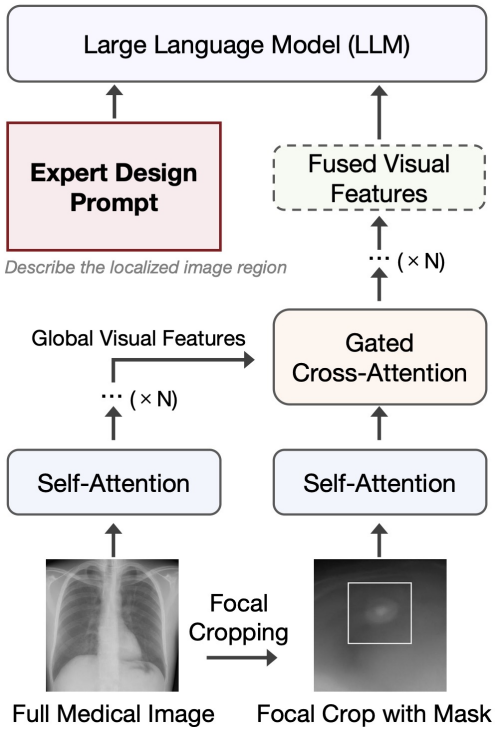

Figure 1. **Architecture of MedDAM.** MedDAM extends the recent breakthrough, i.e., Describe Anything framework (Lian et al., 2025), to medical image understanding. A clinically focused region and its binary mask are used to generate a focal crop, which is embedded along with the full image, fusing global and regional features via gated cross-attention, while structured prompt tokens encode clinical objectives. Then, the resulting features are fed into a LLM to generate region-specific captions.

2. Framework and Evaluation

2.1. Datasets

We evaluate the MedDAM framework (Fig. 1) across three public medical image datasets spanning different imaging modalities and diagnostic contexts. These include chest radiographs, lung CT scans, and dermatological photographs, enabling a comprehensive assessment of the model's generalization on different imaging modalities and organs. The datasets used are summarized in Table 2.

VinDr-CXR (Nguyen et al., 2022) is a large-scale chest X-ray dataset including 18,000 frontal radiographs annotated with bounding boxes and labels for 14 common thoracic abnormalities, such as consolidation, lung opacity, and pleural effusion. It provides a valuable benchmark for evaluating localized captioning in 2D grayscale imaging, where visual abnormalities are often subtle and spatially diffuse.

LIDC-IDRI (Armato III et al., 2011) consists of 1,018 lung CT scans annotated with detailed segmentation masks and semantic attributes for pulmonary nodules, including size, margin sharpness, density, and spiculation. This dataset introduces the challenge of generating localized descriptions within a 3D volumetric image, where fine-grained morphological characteristics play a key role in diagnostic interpretation.

SkinCon (Daneshjou et al., 2022) is a dermatology dataset densely annotated by expert dermatologists, consisting of 3,230 skin lesion images selected from the Fitzpatrick17k dataset, with up to 48 clinical concepts, and reflecting visual characteristics commonly used in clinical skin disease diagnosis.

2.2. Region Sampling and Prompt Construction

Since MedDAM generates descriptions conditioned on localized visual regions, it is important to carefully construct region prompts that reflect clinically meaningful areas while maintaining consistency across datasets. Although the DAM model itself does not rely on manual annotations during inference, we utilize existing annotations (e.g., bounding boxes or segmentation masks), only for evaluation to identify diagnostically relevant regions for prompting. This ensures that the regions of interest used in benchmarking are both clinically grounded and comparable across models. We then standardize the extracted regions (e.g., margin padding, aspect ratio preservation) as input patches to DAM's focal prompting mechanism. In VinDr-CXR, we directly utilize the provided bounding box annotations, each corresponding to a localized thoracic abnormality.

To ensure sufficient contextual information while focusing on the target region, we enlarge each bounding box by a fixed margin of 10%. LIDC-IDRI provides 3D segmentation masks for pulmonary nodules. We extract 2D slices containing nodules and generate bounding boxes from the segmentation masks. For each nodule, we select the axial slice with the most representative appearance (highest radiologist consensus) and crop the minimal enclosed region. For SkinCon, since the dataset doesn't contain any masks or bounding boxes, we implement a region-of-interest detection pipeline adapted from (Wang et al., 2024a) to identify dermatological lesions. Unlike (Wang et al., 2024a), which uses ROI for precise lesion classification, our approach only requires approximate bounding boxes for region-specific captioning purposes. We then extend these detected regions with a slight margin to capture surrounding skin texture variations, providing sufficient context for generating detailed region-specific descriptions. This flexible ROI detection strategy demonstrates how MedDAM can be extended to any medical imaging dataset lacking ground-truth annotations, regardless of imaging modality or anatomical structure, by leveraging existing segmentation techniques to enable region-specific captioning even without manual annotations. Across all the datasets, the extracted region prompts are processed through DAM's focal prompting mechanism, where a high-resolution crop of the region is combined with the full-scale image input. This ensures that the model focuses on the target area while preserving relevant global cues that may influence clinical interpretation. Table 3 summarizes the region sampling strategies and preprocessing steps across the three datasets. To maintain evaluation diversity, we sample up to five regions per image where applicable. For images containing multiple abnormalities, regions are randomly selected to balance anatomical locations and pathology types. Regions are resized to a standard input size while preserving aspect ratio, ensuring consistent processing across all the datasets.

2.3. Attribute Question Construction

Medical image datasets typically lack region-level captions, making traditional text-based evaluation metrics infeasible for our task. Instead, we adopt an attribute-level verification strategy inspired by DLC-Bench (Lian et al., 2025), circumventing the need for ground-truth descriptions as references. For each sampled region, we design a set of clinically relevant binary (yes/no) questions that test whether the generated description correctly reflects the specific attributes of the localized abnormality. The questions are constructed based on the available annotations and the semantic characteristics of each dataset. Positive questions verify whether expected findings are correctly described, while negative questions ensure that unrelated or hallucinated attributes are not mentioned. For VinDr-CXR, questions focus on the presence or absence of radiological features such as consolidation, lung opacity, or pleural effusion within the annotated bounding box. For LIDC-IDRI, questions center around pulmonary nodule properties, including margin sharpness (smooth vs. spiculated), internal density (solid vs. non-solid), and size-related descriptors. For SkinCon, since the original dataset lacks region-level annotations, we generate bounding boxes ourselves using a lightweight lesion detection method (Wang et al., 2024a) as part of our flexible region-of-interest (ROI) pipeline. Based on these detected regions, we construct verification questions focused on key dermatological attributes such as lesion shape, color, border regularity, and surface texture, enabling clinically meaningful evaluation of region-level descriptions. These attributes are commonly used in clinical dermatology for diagnosis, thus serving as reliable semantic anchors for factual verification. Table 4 summarizes the attribute categories and corresponding evaluation focus for each dataset.

2.4. Region-specific Prompting

A key component of realizing MedDAM is the task-specific prompting scheme, namely MedDAM-prompt, tailored to obtaining region-specific captions leveraging pretrained large VLMs. Unlike general captioning prompts, MedDAM-prompt explicitly instructs the model to (i) focus only on the annotated region, (ii) use anatomically precise terminology, and (iii) follow professional clinical report style while avoiding speculative or irrelevant content. As shown in Fig. 2, the prompt includes explicit instructions that define the output format and customized objective. It is designed to minimize hallucinations and ensure the output's high relevance to the localized region.

2.5. Main Results

In Table 1, we report the performance of general-purpose (but adaptable) and region-specific large VLMs on the proposed task. Evaluation metrics include the LLM-score,

Table 1. Main results on LLM-score and MedDLC-score. All models are evaluated in a zero-shot fashion. **Notably, following the practice in (Lian et al., 2025), each score obtained from a model is the average result across the three datasets used in the experiment.**

Model	Type	LLM-score (↑)	MedDLC-score (↑)	Pos QA (↑)	Neg QA (↑)
GPT-4o (OpenAI, 2024)	General	**81.5**	50.2	52.0	48.4
Claude 3.7 Sonnet (Anthropic, 2025)	General	79.2	47.5	49.0	46.0
LLaMA-3.2 Vision (MetaAI)	General	75.3	43.8	45.1	42.5
Qwen2.5-VL (Bai et al., 2025)	General	73.4	41.9	43.2	40.6
GPT-4RoI (Zhang et al., 2023)	Region-specific	77.1	45.7	47.3	44.0
OMG-LLaVA (Zhang et al., 2024)	Region-specific	76.5	46.1	47.8	44.5
MedDAM (Ours)	Region-specific	78.9	**63.6**	**65.1**	**62.0**

which assesses overall language quality, and our MedDLC-score, which reflects clinically grounded, region-specific captioning performance. The latter is further decomposed into positive and negative verification accuracy to provide a more granular analysis. Our MedDAM achieves the highest score on MedDLC-score (i.e., 63.6%), significantly outperforming general models like GPT-4o (i.e., 50.2%) and Claude 3.7 Sonnet (i.e., 47.5%). These results underscore the effectiveness of the region-specific prompting scheme (Sec. 2.4) and the benefit of adapting general localized captioning to medical images. Notably, MedDAM excels in both positive (i.e., 65.1%) and negative (i.e., 62.0%) question types, indicating its ability to not only identify salient findings but also avoid hallucinating ungrounded content—a frequent failure case in general large VLMs. Although MedDAM shows a comprehensive advantage over the baseline models, it trails GPT-4o in terms of LLM-score (i.e., 78.9% vs. 81.5%). It is not surprising since this score is calculated based on a powerful LLM as a judge, and in this experiment we leverage GPT-4o itself as the judge to evaluate its own performance, inevitably incurring bias. In addition to demonstrating the superiority, the results also reveal that MedDAM is capable of generating clinically meaningful and precise descriptions for localized regions, partially facilitated by the proposed evaluation protocol (i.e., MedDLC-score), which uniquely involves clinical significance in evaluating region-specific captioning for medical images. Moreover, MedDAM is free of any ground-truth region-text information as reference. Such a property is highly desired in medical domain, where data annotation could be extremely expensive.

MedDAM-prompt Template

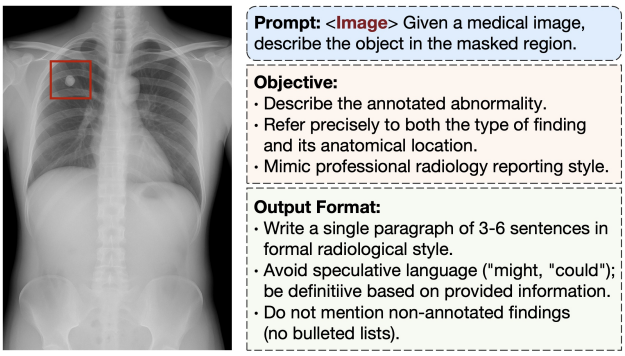

Figure 2. **MedDAM-prompt Template.** This prompt guides the model to produce region-specific, clinically accurate descriptions by incorporating task constraints such as anatomical focus, output format, and information grounding. It is essential for adapting general captioning models like DAM to medical images.

3. Conclusion and Future Work

In this work, we present **MedDAM**, the first framework leveraging the most recent breakthrough, namely Describe Anything Model (DAM), for region-specific captioning in medical images. We show that key to the realization MedDAM includes a proper design of text prompts by medical experts, and an evaluation benchmark to logically assess MedDAM and its competitors. To establish this benchmark, we develop an evaluation protocol tailored to medical images, a data pre-processing pipeline to capture region-level cues from images, and a QA template library, jointly evaluate the performance without resorting to ground-truth descriptions that are scarce in medical images. Our experiments on three publicly available datasets—VinDr-CXR, LIDC-IDRI, and SkinCon—demonstrate that MedDAM significantly improves factual alignment and regional specificity over strong and adaptable large VLMs in zero-shot fashion, also revealing that there is much leeway for boosting the performance of modern VLMs on medical image understanding.

Looking ahead, we envision several promising directions. First, we plan to **fine-tune DAM on large-scale, weakly annotated medical datasets** using pseudo-labeling or self-supervised alignment schemes to further improve the accuracy of generated region-specific captions. Second, we plan to extend MedDAM to additional modalities and specialties, including ophthalmology and pathology, enabling broader benchmarking across various medical image understanding tasks. Finally, it would be interesting to integrate structured knowledge (e.g., RadLex (Radiological Society of North

America (RSNA)), SNOMED (SNOMED International)) into the prompting and evaluation schemes to enhance interpretability and domain alignment of the MedDAM. We expect that this work will inspire more future research on region-specific medical image understanding to eventually benefit clinical applications.

Acknowledgment

This manuscript has been co-authored by ORNL, operated by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy.

Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

Anthropic. Claude 3.7 sonnet. https://www.anthropic.com/claude/sonnet, 2025.

Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Chen, Z., Song, Y., Chang, T.-H., and Wan, X. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.

Daneshjou, R., Yuksekgonul, M., Cai, Z. R., Novoa, R., and Zou, J. Y. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. *Advances in Neural Information Processing Systems*, 35:18157–18167, 2022.

Huang, K.-H., Chen, C.-Y., Chang, K.-W., and Lin, Y.-Y. Radclip: A radiology-specific vision-language foundation model based on clip. *arXiv preprint arXiv:2303.05337*, 2023.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., 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, pp. 590–597, 2019.

Jain, A., Swerdlow, A., Wang, Y., Arnaud, S., Martin, A., Sax, A., Meier, F., and Fragkiadaki, K. Unifying 2d and 3d vision-language understanding. *arXiv preprint arXiv:2503.10745*, 2025.

Jing, B., Xie, P., and Xing, E. Automatic generation of radiology reports: A survey. *arXiv preprint arXiv:1811.02709*, 2018.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Johnson, J., Karpathy, A., and Fei-Fei, L. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016.

Li, J., Li, D., Xiong, C., and Hoi, S. C. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Li, J., Hu, H., Shen, X., Xu, Y., Liu, Z., Xiong, C., and Hoi, S. C. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICLR*, 2023.

Lian, L., Ding, Y., Ge, Y., Liu, S., Mao, H., Li, B., Pavone, M., Liu, M.-Y., Darrell, T., Yala, A., and Cui, Y. Describe anything: Detailed localized image and video captioning, 2025. URL https://arxiv.org/abs/2504.16072.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Liu, S., Zeng, Z., Ren, T., Huang, R., Wang, Y., Xu, H., Li, Z., and Zhang, L. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Ma, J., Wang, H., and et al. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.

Ma, J., Zhang, Y., Wang, H., Wang, Y., et al. Medsam2: Universal medical image segmentation via decomposition and integration. *arXiv preprint arXiv:2403.01928*, 2024.

MetaAI. Llama 3.2: Revolutionizing edge ai and vision with open. Technical report, 2024.

Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., Le, D. D., Pham, C. M., Tong, H. T., Dinh, D. H., et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Radiological Society of North America (RSNA). Radlex radiology lexicon. https://www.rsna.org/practice-tools/data-tools-and-standards/radlex-radiology-lexicon. Accessed: 2025-05-09.

SNOMED International. Snomed ct. https://www.snomed.org/what-is-snomed-ct. Accessed: 2025-05-09.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. *CVPR*, 2015.

Wang, J., Zhang, Y., Ding, Z., and Hamm, J. Achieving reliable and fair skin lesion diagnosis via unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5157–5166, 2024a.

Wang, W., Xiao, X., Liu, M., Lan, Q., Huang, X., Tian, Q., Roy, S. K., and Wang, T. Multi-dimension transformer with attention-based filtering for medical image segmentation. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 632–639. IEEE, 2024b.

Wang, X., Yu, L., Xie, Y., and Xing, L. Medclip: Contrastive learning for medical vision-language processing. *arXiv preprint arXiv:2212.02441*, 2022.

Xiao, X., Wang, W., Xie, J., Zhu, L., Chen, G., Li, Z., Wang, T., and Xu, M. Hgtdp-dta: Hybrid graph-transformer with dynamic prompt for drug-target binding affinity prediction. *arXiv preprint arXiv:2406.17697*, 2024.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5579–5588, 2021.

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

Zhang, T., Li, X., Fei, H., Yuan, H., Wu, S., Ji, S., Loy, C. C., and Yan, S. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *arXiv preprint arXiv:2406.19389*, 2024.

Describe Anything in Medical Images
Supplementary Materials

A. Related Works

A.1. Localized Image Captioning

Image captioning traditionally focuses on generating a global description that summarizes the salient content of an entire image (Vinyals et al., 2015; Anderson et al., 2018). However, many applications demand fine-grained understanding of localized regions rather than holistic summaries. Early works such as DenseCap (Johnson et al., 2016) pioneered dense captioning, which jointly detects regions and generates region-level descriptions. Despite its impact, DenseCap relied heavily on supervised region proposals and densely annotated datasets, limiting its applicability in real-world scenarios. Recent advances in vision-language pretraining have improved the ability to align local visual features with textual representations. VinVL (Zhang et al., 2021) boosted region-level captioning performance by strengthening object detectors and leveraging larger and higher-quality training corpora. BLIP (Li et al., 2022) and BLIP-2 (Li et al., 2023) further advanced this field by using bootstrapped captions and lightweight bridging modules to enhance region-conditioned text generation without full supervision. Grounding DINO (Liu et al., 2023) proposed end-to-end regional grounding by tightly coupling language prompts with visual object detection, enabling open-vocabulary and open-region localization. Nevertheless, these methods often rely on synthetic captions or global image-text pairs, significantly limiting their applications in medical image scenarios. A very recent advance, namely Describe Anything Model (DAM) (Lian et al., 2025), addresses this gap by introducing focal prompting and a localized backbone design, along with a self-supervised data pipeline (DLC-SDP) that automatically constructs detailed region-level pseudo-captions. DAM achieves state-of-the-art performance on localized captioning tasks without requiring region-text annotations, setting a new standard for detailed and scalable region understanding. Despite promising, localized captioning has been predominantly explored in natural images. Extending such models to specialized domains, such as medical image analysis, remains challenging. In this work, we take the first step toward adapting localized captioning models (i.e., originally developed for natural images) to medical images.

A.2. Medical Image Captioning

Medical image captioning aims to automatically generate (e.g., radiology) reports or textual descriptions from medical images. Earlier approaches (Jing et al., 2018; Chen et al., 2020) framed this task as an image-to-sequence problem, applying encoder-decoder architectures. Models such as R2Gen (Chen et al., 2020) introduced memory-driven decoding strategies, while others incorporated hierarchical structures to better reflect the nature of the generated reports. With the availability of large-scale datasets such as CheXpert (Irvin et al., 2019) and MIMIC-CXR (Johnson et al., 2019), supervised report generation has become the major paradigm. More recently, contrastive vision-language pretraining has been adopted to align medical images with their associated reports. MedCLIP (Wang et al., 2022) adapts the CLIP framework (Radford et al., 2021) to report generation through knowledge-driven semantic objectives, while RadCLIP (Huang et al., 2023) introduces volumetric alignment techniques for 2D and 3D medical images. In parallel, segmentation foundation models such as MedSAM (Ma et al., 2023) and MedSAM2 (Ma et al., 2024) have enabled universal promptable segmentation across diverse imaging modalities, focusing on delineating anatomical structures and lesions, however, these models generate masks rather than textual descriptions, leaving the semantic interpretation of regions to human users. While report generation models capture global image-level information, and segmentation models extract structural regions, the intermediate task of producing localized and detailed textual descriptions of specific regions remains underexplored. Recent efforts such as UniVLG (Jain et al., 2025) have unified multiple vision-language tasks, but still operate mainly at image level. To date, there exists no systematic study of localized region-level captioning for medical images. *Our work fills this gap by realizing a deployable framework MedDAM (Fig. 1) to facilitate the zero-shot transfer of natural-image-based localized captioning models, including the very recent breakthrough, namely DAM (Lian et al., 2025), to medical image understanding.*

B. Baselines and Evaluation Metrics

We compare MedDAM against a series of state-of-the-art adaptable large VLMs that can understand visual content via texts. Specifically, we evaluate GPT-4o, Claude 3.7 Sonnet, LLaMA-3.2 Vision, Qwen2.5-VL, GPT-4Rol and

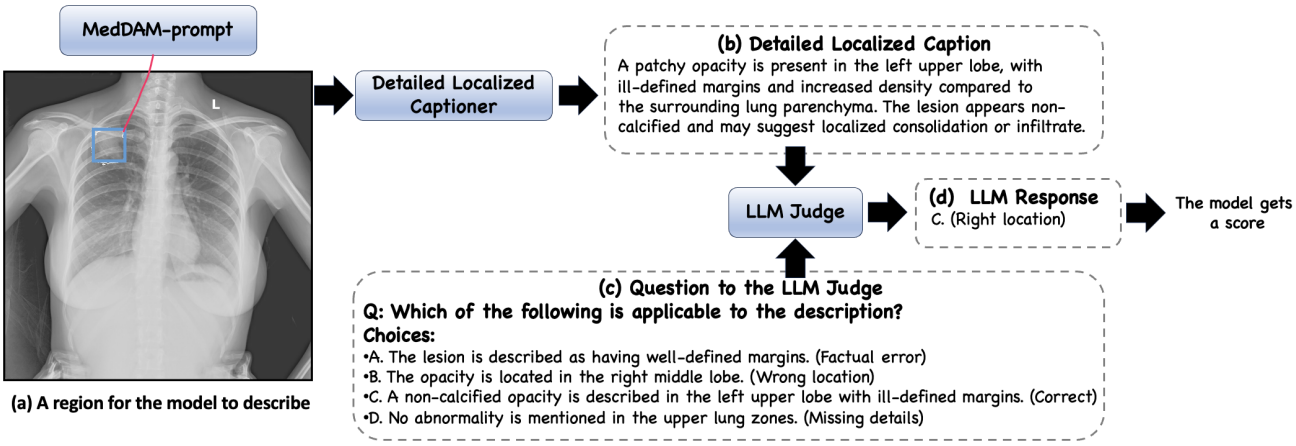

Figure 3. **An example evaluation pipeline (i.e., calculating MedDLC-score).** (a) A region of interest is marked in a chest X-ray image and then used as input to MedDAM, prompted by a task-specific instruction. (b) The model generates a region-specific caption describing the abnormality within the marked region. (c) A question-answering task is set up to verify the factual accuracy and localization consistency of the generated caption, based on domain-specific attributes. (d) An LLM-based evaluator assigns a correctness label to the answer, and the model receives a score if the description matches the ground-truth semantic attributes. This framework enables reference-free benchmarking of fine-grained regional captioning performance on medical image understanding. **Although both the widely used LLM-score and our MedDLC-score involve a LLM Judge, the former is more effective for natural vision-language scenarios while the latter is tailored to the proposed task in medical domain.**

OMG-LLaVA. These models represent the current frontier of vision-language pretraining and instruction tuning, covering a diverse range of architectures and training paradigms. For all the baselines, we provide region-level crops as inputs and retain access to the full-scale image input, as we do for MedDAM, prompting the models to generate region-specific descriptions. An example evaluation pipeline is illustrated in Fig. 3. To measure model performance, we adopt two protocols. First, we use the LLM-score, in which an independent and strong LLM (i.e., GPT-4o) serves as a judge to assess the fluency, relevance, factual correctness, and clinical plausibility of the generated descriptions. Each factor is rated individually, and the final score is computed by averaging across all evaluated regions. Second, we propose a clinically grounded and reference-free evaluation protocol, namely MedDLC-score, tailored to medical semantic attributes. Instead of relying on ground-truth captions, it formulates a set of attribute-level binary verification tasks, assessing whether the generated descriptions accurately capture key clinical features such as lesion location, morphological appearance, and radiological findings. Model performance is reported as accuracy over positive and negative verification questions. This dual evaluation strategy facilitates a comprehensive assessment of both linguistic quality and medical factuality, yielding valuable feedback on the capabilities and limitations of region-specific captioning models in medical image understanding.

Table 2. Summary of the datasets used for evaluating MedDAM.

Dataset	Modality	Region Annotation (Source & Type)	Region Sampling Strategy & Evaluation Focus	Sample Size
VinDr-CXR (Nguyen et al., 2022)	Chest X-ray (CXR)	Bounding boxes (thoracic abnormalities); 2D bounding boxes	Use all annotated boxes with 10% margin expansion; evaluate lesion localization, opacity patterns, consolidation, and effusion findings	18,000 images
LIDC-IDRI (Armato III et al., 2011)	Lung CT scan (3D)	Segmentation masks (nodules); 2D bounding boxes from selected slices	Select slice with highest radiologist agreement; extract tight bounding boxes from 2D mask slices; evaluate nodule morphology including margin, spiculation, and internal texture	1,018 scans
SkinCon (Daneshjou et al., 2022)	Clinical photography	Bounding boxes (cutaneous lesions); 2D bounding boxes	Use all annotated lesion regions with margin padding; evaluate dermatological attributes such as shape, color, border regularity, and surface texture	3,000 images

Table 3. Summary of the region sampling and prompt construction (Sec. 2.2) for each dataset.

Dataset	Region Source and Type	Selection Strategy	Margin Handling and Processing	Notes
VinDr-CXR (Nguyen et al., 2022)	Bounding boxes (thoracic abnormalities); 2D crops	Use all annotated bounding boxes with small context preservation	Expand bounding box by 10%; resize crop to fixed input size while keeping aspect ratio	Focus on capturing localized opacities and consolidations within thoracic regions
LIDC-IDRI (Armato III et al., 2011)	3D segmentation masks (lung nodules); 2D slices from mask	Select axial slice with highest radiologist agreement; extract tight bounding box around mask	No margin expansion; resize crop to fixed input size while keeping aspect ratio	Emphasizes fine-grained pulmonary nodule features such as spiculation, density variations, and margin sharpness
SkinCon (Daneshjou et al., 2022)	Bounding boxes (cutaneous lesions); 2D crops from photographs	Use all visible lesion annotations; preserve color texture cues from skin	Expand bounding box by 15% for context; resize crop to fixed size while preserving RGB fidelity	Prioritizes dermatological features such as lesion shape, color, boundary regularity, and skin texture context

Table 4. Summary of the attribute verification tasks (Sec. 2.3).

Dataset	Attribute Type	Positive QA Example	Negative QA Example
VinDr-CXR (Nguyen et al., 2022)	Radiological findings (opacity, consolidation, effusion)	Is there increased opacity in the lower lobe of the lungs?	Is pneumothorax incorrectly mentioned in the localized description?
LIDC-IDRI (Armato III et al., 2011)	Pulmonary nodule morphology (margin, density, size)	Does the description mention a spiculated nodule margin?	Is lobulation falsely described when it is not present?
SkinCon (Daneshjou et al., 2022)	Dermatological lesion characteristics (color, shape, texture)	Does the caption describe an irregular border and reddish hue?	Is scaling or ulceration incorrectly attributed to the lesion?