PathVG: A New Benchmark and Dataset for Pathology Visual Grounding

Chunlin Zhong[1⋆] Shuang Hao[1⋆], Junhua Wu[2⋆], Xiaona Chang[2⋆], Jiwei Jiang[1], Xiu Nie[2✉], He Tang[1✉], Xiang Bai[1]

[1] School of Software Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[2] Department of Pathology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China
{clzhong,hetang}@hust.edu.cn, niexiuyishi@126.com

Abstract. With the rapid development of computational pathology, many AI-assisted diagnostic tasks have emerged. Cellular nuclei segmentation can segment various types of cells for downstream analysis, but it relies on predefined categories and lacks flexibility. Moreover, pathology visual question answering can perform image-level understanding but lacks region-level detection capability. To address this, we propose a new benchmark called Pathology Visual Grounding (PathVG), which aims to detect regions based on expressions with different attributes. To evaluate PathVG, we create a new dataset named RefPath which contains 27,610 images with 33,500 language-grounded boxes. Compared to visual grounding in other domains, PathVG presents pathological images at multi-scale and contains expressions with pathological knowledge. In the experimental study, we found that the biggest challenge was the implicit information underlying the pathological expressions. Based on this, we proposed Pathology Knowledge-enhanced Network (PKNet) as the baseline model for PathVG. PKNet leverages the knowledge-enhancement capabilities of Large Language Models (LLMs) to convert pathological terms with implicit information into explicit visual features, and fuses knowledge features with expression features through the designed Knowledge Fusion Module (KFM). The proposed method achieves state-of-the-art performance on the PathVG benchmark. The source code and dataset have been available at https://github.com/ssecv/PathVG.

Keywords: Pathology Visual Grounding · Vision-Language Model · Large Language Model.

1 Introduction

Pathology is the cornerstone of modern medicine, playing a crucial role in disease diagnosis and understanding. With the development of artificial intelligence, computational pathology has made significant strides, such as whole-slide cancer subtyping and survival prediction [7, 9, 11], cellular nuclei segmentation [16],

⋆ Equal contribution ✉ Corresponding author

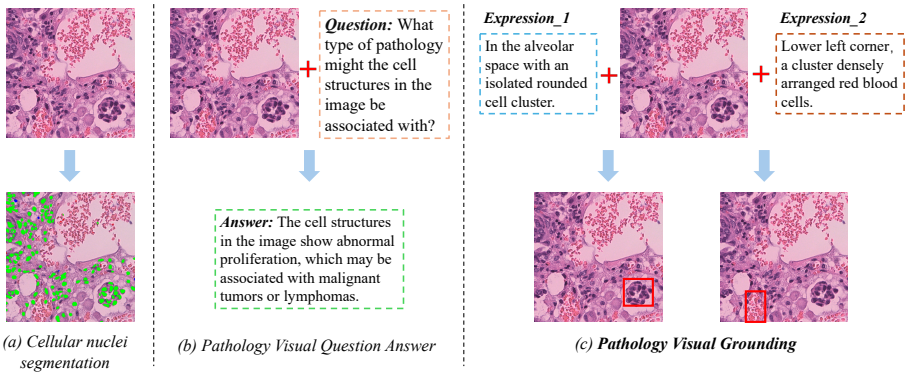


Fig. 1: A comparison of (a) Cellular nuclei segmentation, (b) Pathology Visual Question Answer and (c) our proposed PathVG benchmark.

and pathology visual question answering [13, 5]. However, current computational pathology tasks still face certain limitations. Cancer subtype and survival prediction are based on overall predictions from whole slide images, and patch-level analysis cannot be performed. Cellular nuclei segmentation (Fig. 1(a)) can segment various types of cells for downstream analysis, but it relies on predefined categories and lacks flexibility. Moreover, pathology visual question answering (Fig. 1(b)) focuses on image-level understanding and cannot perform region-level detection. In clinical practice, however, it is often necessary to detect different regions based on factors such as different organs and cancer types or in response to referring human input. To address these challenges, we propose a novel benchmark, PathVG, which provides flexible and region-level detection capabilities. In contrast, PathVG(Fig. 1(c)) allows for the localization of different regions by inputting various expressions.

In recent years, Medical Visual Grounding (MVG) has already been explored [1, 2, 10, 6]. As shown in Fig.2 (a) and (b), due to the uniqueness of pathological data, PathVG presents two key distinctions compared to the previous MVG: (1) **Multi-scale pathological images:** A uniqueness of pathological images is that the same region exhibits different pathological features at varying magnification levels. Higher magnification images highlight cell structure and growth, while lower-magnification images reveal cell arrangement and interaction with neighboring cells. (2) **Expressions with pathological knowledge**: PathVG localizes specific regions from multiple pathological perspectives, such as cell structure, growth patterns, cell arrangement, and interactions with neighboring cells, to accurately localize regions under different magnification. To align with these two distinctions, we introduce a novel dataset called RefPath, specifically tailored for PathVG. PathVG includes 27,610 images with 33,500 language-grounded boxes.

Compared to visual grounding in other domains, the main challenge of PathVG lies in the implicit information underlying the pathological expressions, which

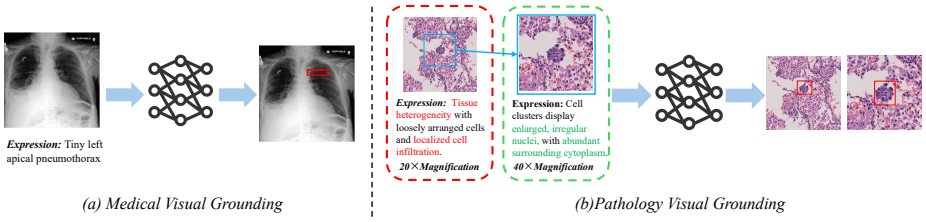


Fig. 2: (a) Previous Medical Visual Grounding. (b) Pathology Visual Grounding: Identical region at Lower **(Left)** and Higher **(Right)** Magnification, with expressions for cell arrangement and interactions with neighboring cells (**Red**), as well as cell structure and growth (**Green**).

makes it difficult to associate them with pathological images. The expressions in RefPath describe pathological region from multiple perspectives, requiring the model to understand a wide range of specialized terms. This becomes a challenging task for models without prior knowledge of pathology.

To address this challenge, we introduced the Pathology Knowledge-enhanced Network (PKNet), which leverages the knowledge enhancement capabilities of LLMs, transforming implicit pathological terms into explicit visual information, and better linking pathological expressions with pathological images. Building upon this, we have designed a Knowledge Branch specifically for knowledge enhancement, as well as a Knowledge Fusion module (KFM) to better fuse knowledge and expression features.

In summary, our contributions are listed as follows:

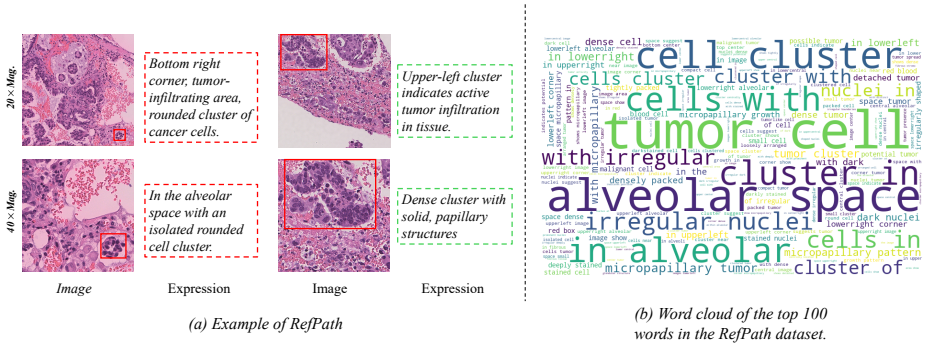
1. We propose a novel benchmark, Pathology Visual Grounding (PathVG), which enables flexible and region-level detection in pathological images.
2. We present RefPath, a large-scale dataset consisting of 27,610 images and 33,500 language-grounded boxes, tailored to the uniqueness of pathology.
3. We developed a baseline model, the Pathology Knowledge-enhanced Network (PKNet), which leverages knowledge enhancement from LLMs to transform implicit pathological expressions into explicit visual features.

2 RefPath: A Large-scale Dataset for PathVG

2.1 Dataset Collection and Annotation

To adapt to the PathVG benchmark, we have specifically built a new large-scale dataset, RefPath. To ensure the quality of our dataset, we have carefully devised a three-step data processing and generation protocol.

Step 1: Data Collection and Preprocessing. The first step involved collecting clinical whole-slide images, cropped at 40× and 20× magnifications with a resolution of 1024×1024 pixels. RefPath was obtained from our collaborating hospital and approved by its institutional ethics committee. 40× magnification



(a) Example of RefPath

(b) Word cloud of the top 100 words in the RefPath dataset.

Fig. 3: (a): Example of RefPath: Displayed the differences in images and at different magnification levels. (b): Word cloud of the top 100 words in the RefPath dataset. Displayed the pathological terms in the expression.

is used to examine fine-grained cellular morphology, such as nuclear features, staining patterns, and structural abnormalities. In contrast, 20× magnification assesses broader tissue context, including cell cluster arrangement, spatial distribution, and glandular organization. Pathological experts manually labeled language-grounded boxes to identify cell clusters or regions indicative of malignancy. These annotations trained the YOLOv10 [15] network for automated region detection. Candidate boxes were reviewed by experts to filter out irrelevant or unclear annotations, ensuring only high-quality data for downstream steps.

Step 2: Pathological expression Generation. In addition to language-grounded boxes annotations, pathological experts provided detailed expressions for selected images. These expression annotations served as key examples for few-shot learning by GPT-4V, improving its ability to generate accurate, context-rich pathological expressions. To optimize GPT-4V's output, we designed specific prompts to focus the model on the morphological features of cells and tissues within the annotated regions, ensuring high precision and relevance.

Step 3: Expert Validation. We divided the dataset into training and testing sets, and invited professional pathology experts to manually review the testing set. Five attending-level or above pathologists joined in the annotation process. A standardized annotation protocol was jointly defined, and a cross-checking strategy was used to ensure consistency. The experts evaluated the textual expression based on the following criteria: (1) whether they conform to standard clinical expression; (2) whether they correspond accurately to the specified regions; and (3) whether they provide a multifaceted, fine-grained portrayal of the specified regions. Samples failing to meet any of these criteria were deemed invalid and removed from the RefPath dataset. Ultimately, the training set comprises 24757 images with 30452 language-grounded boxes, while the testing set includes 2853 images with 3048 language-grounded boxes.

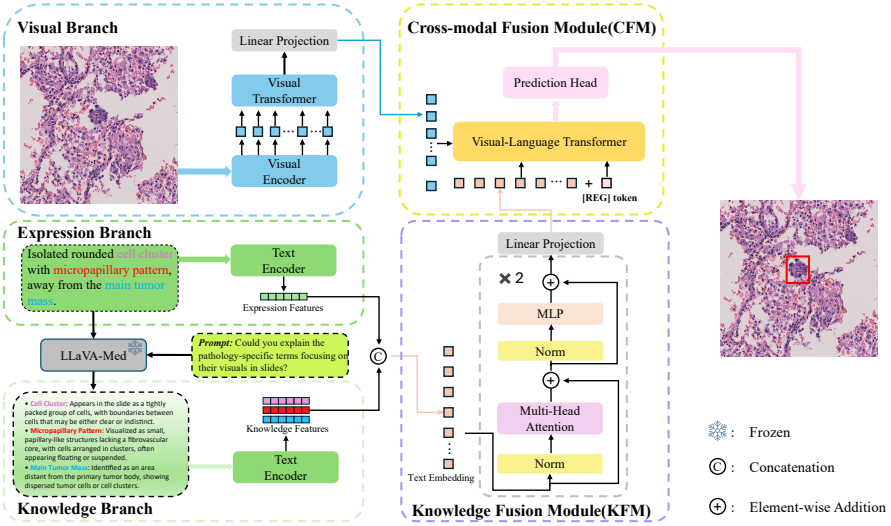


Fig. 4: Overview of the proposed method. The model uses the knowledge enhancement ability of LLMs to connect pathological expressions with visual features, integrates Expression and Knowledge features through KFM, and outputs the final language-grounded boxes by combining visual features with CFM.

2.2 Dataset Statistics

As shown in Fig.3, we present the uniqueness of the Refpath dataset from two aspects. First, the dataset contains multi-scale images, as shown in Fig.3(a). At low(20×) magnification, the images emphasize cell arrangement and interactions with neighboring cells, such as infiltration observed in the expression. At high(40×) magnification, the focus shifts to cell structure and growth, such as alveolar spaces and papillary structures observed in the expression. Secondly, the Expression contains various pathology-related terms, as shown in Fig.3(b). We can observe that the RefPath dataset includes pathological terms such as 'tumor cells,' 'alveolar spaces,' 'irregular nuclei,' and so on.

3 Method

3.1 Model Architecture

As shown in Fig. 4, PKNet consists of five main components:(1) Visual Branch, (2) Expression Branch, (3) Knowledge Branch, (4) Knowledge Fusion Module, and (5) Cross-modal Fusion Module(CFM).

Visual Branch: Following the common practice [14], the visual encoder starts with a CNN backbone, followed by the visual transformer.We choose the ResNet-50 as the CNN backbone. The visual transformer includes 6 stacked transformer encoder layers. Each transformer encoder layer includes a multi-head

self-attention layer and an FFN. Given an pathological image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$ as input of ResNet-50 to generate a 2D feature map $\mathcal{Z} \in \mathbb{R}^{C_v\times H_v\times W_v}$. The channel dimension C_v is 256 and the width W_v and height H_v of the 2D feature map are $\frac{1}{32}$ of the original image size. Next, we flatten $\mathcal{Z}$ into $\mathcal{Z}_v \in \mathbb{R}^{C_v\times N_v}$, where $N_v = H_v * W_v$. Finally, pass $\mathcal{Z}_v$ through the transformer block to obtain the final visual feature $\mathcal{F}_v \in \mathbb{R}^{C_v\times N_v}$.

Expression Branch: We use the 12-layer BERT as our expression text encoder. Given an expression as the input of this branch, we first convert each word ID into a one-hot vector. Then, in the token embedding layer, we tokenize each one-hot vector into a language token. After that, we take the language tokens as inputs of the expression transformer, and generate the advanced language features $\mathcal{F}_e \in \mathbb{R}^{C_e\times N_e}$, where C_e is the output dimension of transformer, N_e is the number of language tokens. The process is formulated as follows:

$$\mathcal{F}_e = \mathcal{E}_t(T, \theta_t), \tag{1}$$

where $\mathcal{E}_t(\cdot, \theta_t)$ is the text encoder stated above. θ_t denotes the encoder parameters.

Knowledge Branch: The Knowledge Branch is similar to the Expression Branch. Considering the similarity between knowledge and expression, we use the same encoder to extract knowledge features $\mathcal{F}_k$. The process is formulated as follows:

$$\mathcal{F}_k = \mathcal{E}_t(\mathcal{H}(T, P), \theta_t), \tag{2}$$

where $\mathcal{H}$ represents the LLM we use to associate pathological terms with corresponding visual features, and P is the prompt used in the LLM.

Knowledge Fusion Module: After the individual expression and knowledge encoding, we obtain $\mathcal{F}_e$ and $\mathcal{F}_k$. To integrate these two features, we propose KFM, which consists of two transformer layers. Each layer includes a multi-head self-attention layer(MSA) and a FFN. Use KFM to get the language features $\mathcal{F}_l$. The process of KFM is formulated as follows:

$$\mathcal{F}_l = FFN(MSA(Concat(\mathcal{F}_e, \mathcal{F}_k)). \tag{3}$$

Cross-modal Fusion Module: The CFM module includes two linear projection layers (one for each modality) and a visual-language(V-L) transformer (with a stack of 6 transformer encoder layers). CFM is used to integrate the fused $\mathcal{F}_v \in \mathbb{R}^{C_v\times N_v}$ with $\mathcal{F}_l \in \mathbb{R}^{C_l\times N_l}$. First, the features of both modalities project to the same channel dimension through a linear projection layer. We denote the projected visual features and textual features as $\mathcal{P}_v \in \mathbb{R}^{C_p\times N_v}$ and $\mathcal{P}_l \in \mathbb{R}^{C_p\times N_l}$, respectively. Then, we prepend a learnable embedding ([REG] token) join to $\mathcal{P}_v \in \mathbb{R}^{C_p\times N_v}$ and $\mathcal{P}_l \in \mathbb{R}^{C_p\times N_l}$ as the input $\mathcal{X}_0$. After that, we input $\mathcal{X}_0$ into the V-L transformer to obtain REG_{output}. Finally, we leverage REG_{output} from the V-L transformer as the input of our prediction head. The prediction head consists of an MLP layer. The output of it is a sequence b organized as (x, y, w, h) that means the coordinates of the top left vertex, the width, and the length for the regressed bounding box.

Table 1: PathVG results on RefPath with respect to Acc and mIoU. ↑ denotes that a larger value is better. We highlight the best in the **red**. † represents Multimodal Large Language Model.

Model	Venue	Visual/Text Encoder	$RefPath_{all}$		$RefPath_{40\times}$		$RefPath_{20\times}$	
			$Acc\uparrow$	$mIOU\uparrow$	$Acc\uparrow$	$mIOU\uparrow$	$Acc\uparrow$	$mIOU\uparrow$
TransVG[4]	ICCV'21	RN50/BERT-B	58.40	52.86	68.72	66.75	50.29	41.94
SeqTR[18]	ECCV'22	DN53/BiGRU	55.84	51.96	72.65	71.13	42.57	36.78
CLIPVG[17]	TMM'23	CLIP-B/CLIP-B	58.89	53.97	75.52	72.14	45.81	39.67
LLaVa-Med†[8]	NeuIPS'23	CLIP-L/LLaMa	62.32	57.96	73.52	70.24	53.51	48.31
TransCP[14]	TPAMI'24	RN50/BERT-B	61.73	56.81	74.27	71.92	51.87	44.93
SimVG[3]	NeurIPS'24	ViT-B/BERT-B	63.94	59.42	75.36	73.18	52.52	46.92
D-MDETR [12]	TPAMI'24	CLIP-B/CLIP-B	64.92	57.69	76.29	73.10	55.98	45.57
Ours	-	RN50/BERT-B	**69.95**	**63.49**	**80.48**	**76.88**	**61.66**	**52.95**

3.2 Loss Function

The model's training uses the L1 and IoU loss functions.

$$\mathcal{L} = \lambda_{l1}\mathcal{L}_{l1}(\mathcal{P}, \mathcal{GT}) + \lambda_{iou}\mathcal{L}_{iou}(\mathcal{P}, \mathcal{GT}), \tag{4}$$

where $\mathcal{P}$ and $\mathcal{GT}$ denote the regressed bounding box and the ground truth bounding box, respectively. λ_{l1} and λ_{iou} are two trade-off factors that balance the two losses which are set to 5 and 2 empirically.

4 Experiment

Dataset. The RefPath dataset includes 27,610 images with 33,500 language-grounded boxes. The training set has 24,757 images with 30452 language-grounded boxes, while the test set contains 2,853 images with 3,048 language-grounded boxes. It is further divided into two subsets: the 40× subset with 1,342 language-grounded boxes and the 20× subset with 1,706 language-grounded boxes. All methods are evaluated on the same training and test sets.

Evaluation Metric. To evaluate the model's performance on the PathVG, we follow the standard protocol for visual grounding [4] to report accuracy (Acc%). Due to the unique nature of pathological images at different magnifications, using the same IoU threshold across magnifications is unreasonable. Therefore, we set the IoU threshold to 0.7 for 40× images and 0.5 for 20× images. Additionally, we use mIoU% for a more comprehensive comparison.

Implementation details. We use a single NVIDIA GeForce RTX 3090 GPU for training and testing. The weights of the CNN backbone and Transformer encoder are initialized using the pre-trained DETR model. The AdamW optimizer is employed with an initial learning rate of 1e-5. The model is trained for a total of 90 epochs. For the other comparative methods, we follow the training and testing configurations specified in their respective papers. *Our base model*

Table 2: Ablation Study. (b) refers to joining the Knowledge text and expression text into a long text, which is then input into the Expression Branch; (c) uses a Knowledge Branch to extract knowledge features; (d) builds upon (c) by adding the KFM module to fuse knowledge features and expression features.

	Know. Input	Know. Branch	KFM	$RefPath_{all}$		$RefPath_{40\times}$		$RefPath_{20\times}$	
				$Acc\uparrow$	$mIOU\uparrow$	$Acc\uparrow$	$mIOU\uparrow$	$Acc\uparrow$	$mIOU\uparrow$
(a)				61.73	56.81	74.27	71.92	51.87	44.93
(b)	✓			62.29	58.38	74.65	73.13	52.57	46.78
(c)	✓	✓		67.60	61.03	77.69	74.34	59.67	50.56
(d)	✓	✓	✓	**69.95**	**63.49**	**80.48**	**76.88**	**61.66**	**52.95**

does not use a pre-aligned vision-language model, and all the pre-trained vision encoder and text encoder have not seen pathological images.

Results on PathVG. Tab. 1 presents a comparison of the performance of different models on the RefPath dataset. The compared methods include TransVG[4], SeqTR[18], CLIPVG[17], LLava-Med[8], TransCP[14], SimVG[3] and Dynamic-MDETR[12]. For LLaVa-Med, We changed the dataset format to the Med-GRIT-270k[6] dataset to train LLaVa-med. As can be seen, our method achieves the best performance across all evaluation metrics, especially on the more challenging $RefPath_{20\times}$. This outstanding performance can be attributed to the need for more accurate understanding of the pathological expression in lower-magnification images. In contrast, SeqTR performs poorly primarily because its text encoder is a simple BiGRU, which limits its ability to comprehend pathological expression.

Ablation Study. We conducted ablation experiments on the Refpath to evaluate the effectiveness of each component in PKNet. Tab. 2 presents the quantitative results for each configuration. First, in the baseline setup without additional information input, as shown in (a) in Tab. 2, the performance is suboptimal. Next, we consider incorporating knowledge information as an additional input, as shown in (b). By simply concatenating the two text as input, we observe a slight improvement, though the effect is not significant. Subsequently, to better extract both knowledge and expression information, we designed a dedicated Knowledge Branch, which led to a considerable improvement. Finally, we introduced a specially designed KFM module to fuse knowledge features and expression features, resulting in further gains and achieving the best performance.

5 Conclusion

In this paper, to address the limitations of existing computational pathology tasks. We propose a new benchmark, PathVG, and a dedicated dataset, RefPath. PathVG is a new benchmark that enables precise localization of specific regions in pathological images using fine-grained text descriptions. The proposed RefPath dataset contains over 27,000 images with detailed annotations. Building upon

this, we introduce a new base model, PKNet, which leverages the knowledge-enhancement capabilities of large models to effectively bridge the gap between pathological expression and images.

Limitation. The benchmark and method proposed in this paper are based on fully supervised learning. However, in the medical field, obtaining annotations is costly. Therefore, we plan to explore semi-supervised and unsupervised learning approaches in future work to reduce reliance on expensive annotations. This work focuses on establishing a foundational one-to-one grounding benchmark. We excluded 10× to avoid one-to-many mapping ambiguity. One-to-many grounding and broader magnifications are our future research directions, which will be discussed in the final version.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant 62225603 and 62476106), the National Key R&D Program of China (2022YFF1203300), the Natural Science Foundation of Hubei Province of China (No. 2024AFB545), and the Fundamental Research Funds for the Central Universities (Grand No. YCJJ20252416). We thank Dr. Yang Jin for his helpful comments and his pivotal role in resource coordination. The computation has been completed in the HPC Platform of Huazhong University of Science and Technology.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.
2. Zhihao Chen, Yang Zhou, Anh Tran, Junting Zhao, Liang Wan, Gideon Su Kai Ooi, Lionel Tim-Ee Cheng, Choon Hua Thng, Xinxing Xu, Yong Liu, et al. Medical phrase grounding with region-phrase context contrastive alignment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 371–381. Springer, 2023.
3. Ming Dai, Lingfeng Yang, Yihao Xu, Zhenhua Feng, and Wankou Yang. Simvg: A simple framework for visual grounding with decoupled multi-modal fusion. *Advances in Neural Information Processing Systems*, 37:121670–121698, 2025.
4. Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021.
5. Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.
6. Xiaoshuang Huang, Haifeng Huang, Lingdong Shen, Yehui Yang, Fangxin Shang, Junwei Liu, and Jia Liu. A refer-and-ground multimodal large language model for biomedicine. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 399–409. Springer, 2024.
7. Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.
8. Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.
9. Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.
10. Nevin M Matasyoh, Rüdiger Schmidt, Ramy A Zeineldin, Uwe Spetzger, and Franziska Mathis-Ullrich. Interactive surgical training in neuroendoscopy: Real-time anatomical feature localization using natural language expressions. *IEEE Transactions on Biomedical Engineering*, 2024.
11. Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.
12. Fengyuan Shi, Ruopeng Gao, Weilin Huang, and Limin Wang. Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1181–1198, 2023.
13. Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Dan Wan, Xiaoxiao Lan, Mengyue Zheng, et al. Pathmmu: A massive

multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, pages 56–73. Springer, 2024.
14. Wei Tang, Liang Li, Xuejing Liu, Lu Jin, Jinhui Tang, and Zechao Li. Context disentangling and prototype inheriting for robust visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3213–3229, 2023.
15. Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2025.
16. Huyong Wang, Huisi Wu, and Jing Qin. Incremental nuclei segmentation from histopathological images via future-class awareness and compatibility-inspired distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11408–11417, 2024.
17. Linhui Xiao, Xiaoshan Yang, Fang Peng, Ming Yan, Yaowei Wang, and Changsheng Xu. Clip-vg: Self-paced curriculum adapting of clip for visual grounding. *IEEE Transactions on Multimedia*, 26:4334–4347, 2023.
18. Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022.