PCRP: Progressive Coarse-to-Fine Refinement for Pathology Grounding

Liang Peng[1,*], Chenxiao Li[2,*], Shaohua Dong[2], Bohan Tan[3], Zhipeng Zhang[4], Xingping Dong[1,†]

[1]School of Computer Science, National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University
[2]University of North Texas [3]Huazhong University of Science and Technology
[4]SAI, Shanghai Jiao Tong University

2025102110090@whu.edu.cn, ChenxiaoLi@my.unt.edu, xingpingdong@whu.edu.cn

Abstract. Pathology Visual Grounding aims to precisely localize the pathological tissue region in a histopathology image and output a bounding box based on professional medical language descriptions. However, in histopathology images, the background textures are highly similar and the target regions are small, low in contrast, and have blurred boundaries, and medical descriptions cannot be accurately understood, making it difficult for the described location in the text to correspond to a unique target region. Existing pathology visual grounding methods make insufficient use of fine-grained visual cues in images and rely more on coarse language-vision alignment, which easily causes misalignment and drift on difficult samples such as small targets, occlusion, or blurred boundaries. To address these issues, we propose PCRP (Progressive Coarse-to-Fine), a progressive framework for pathology visual grounding that improves grounding accuracy by strengthening reliable visual cues and performing robust boundary refinement. PCRP consists of four modules,Visual Branch, Text Branch, VWF and C2F-BAR. This design reduces misalignment and drift on samples with small targets and blurred boundaries, thereby improving robustness and grounding quality. PCRP achieves state-of-the-art performance on RefPath ($RefPath_{all}$) and significantly outperforms PKNet by 11.68% Acc and 9.27% mIoU, validating its effectiveness. The source code is available at `https://github.com/wuhengliangliang/PCRP`.

Keywords: Pathology Visual Grounding · SAM3.

1 Introduction

Pathology is the cornerstone of medical diagnosis, where the confirmation of many diseases relies not only on recognizing pathological patterns but also

* Equal contribution. † Corresponding author.

on accurately localizing specific regions. In recent years, computational pathology has achieved remarkable progress in whole-slide classification and prognosis modeling [8,10,12], nucleus or tissue segmentation for quantitative analysis [16], and pathology-oriented visual question answering for semantic interpretation [14,6]. Image-level medical visual grounding (MVG) has also advanced significantly [1,3,7,11]. However, in clinical practice, pathologists often need to precisely localize specific regions in slides based on natural language descriptions. The grounding of regions across multi-scale pathology images and multi-perspective pathological expressions remains underexplored. Therefore, Pathology Visual Grounding (PathVG) emerges as an important yet challenging task to drive fine-grained computational pathology analysis.

Existing approaches leverage the knowledge-enhancement capability of large language models [18] to transform implicit pathological terminology into explicit visual information, thereby bridging the gap between pathological expressions and images. Nevertheless, two major limitations persist: (1) Semantic drift: pathological expressions are usually short and highly specialized. Even with knowledge enhancement, language tokens can still spread across irrelevant context during cross modal alignment, resulting in weak spatial specificity for the referred region, especially in low magnification images with small target regions. (2) Ambiguous visual boundaries: current supervision is mostly box-level annotation, lacking pixel-level boundary information. In complex tissue interfaces or small target regions, predictions often drift and boundaries remain imprecise.

To address these challenges, we propose **P**rogressive **C**ross-modal alignment with **R**eliable mask **P**riors (**PCRP**), a framework combining cue-guided language alignment and reliability-controlled mask priors for boundary refinement. PCRP reduces language-side semantic ambiguity and improves box-supervised boundary grounding for robust pathology visual grounding. Specifically, it builds a **Visual Branch** and a **Text Branch** to extract complementary grounding features. We introduce Visual-Word Fusion (**VWF**), which retrieves word-wise visual evidence via text-to-vision cross-attention (Q=text, K/V=visual) and injects it into text features through a learnable gate, improving low-magnification small-target grounding. To improve grounding precision, we introduce a two-stage regressor **C2F-BAR**: Stage 1 predicts a stable coarse box, and Stage 2 leverages pixel-level masks from teacher SAM3 [2] as boundary priors. With a curriculum-style strategy, the prompt source shifts from ground-truth to predicted boxes for fine-grained refinement. Overall, PCRP alleviates textual semantic drift and visual-anchor boundary ambiguity, enabling robust pathology visual grounding under challenging conditions such as texture homogeneity and tiny targets. Experiments show that PCRP significantly outperforms existing methods on RefPath, with a 16% gain on the 20× low-magnification dataset.

Our contributions can be summarized as follows:

1. We propose **PCRP**, a progressive pathological visual grounding framework that simultaneously alleviates semantic shifts in language and blurring of visual segmentation boundaries.

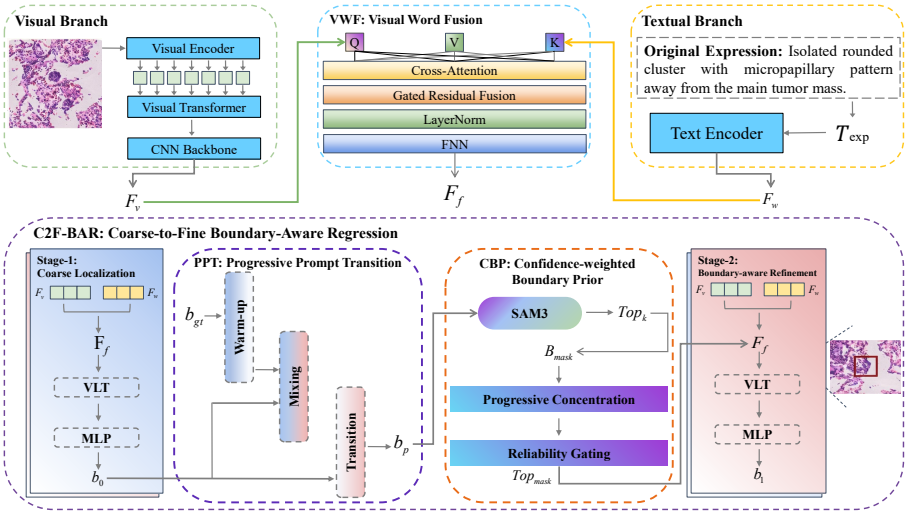

Fig. 1: Overview of our PCRP framework.

2. We propose Visual-Word Fusion (**VWF**) and Coarse-to-Fine Boundary-Aware Regression (**C2F-BAR**) to achieve multimodal semantic alignment and more refined and reliable visual boundaries within specific regions.
3. Extensive experimental results demonstrate that our method achieves state-of-the-art (SOTA) performance on RefPath, particularly for tiny targets and scenes under 20× low magnification.

2 Method

Given a histopathology image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$ and its referring expression T_{exp}, our goal is to localize the target region by predicting a 4-dimensional language-grounded box $b \in [0,1]^4$ in (x, y, w, h) format, where (x, y) denotes the box center and w and h are its width and height, respectively.

2.1 Model Architecture

In this paper, we propose **PCRP** for pathology visual grounding. As shown in Fig. 1, **PCRP** consists of four components, *i.e.*, **Visual Branch**, **Text Branch**, **VWF**, and **C2F-BAR**. VWF progressively injects image-conditioned local visual cues into word-level textual features, improving fine-grained spatial alignment between referential pathology descriptions and target regions, which is especially beneficial for tiny targets under repetitive textures, weak contrast, and adjacent-cell interference. Based on the resulted multimodal representations, C2F-BAR performs two-stage progressive box refinement and incorporates

reliable mask priors for boundary-aware localization, with CBP (Confidence-weighted Boundary) and PPT(Progressive Prompt Transition) jointly controlling prior quality and the stage of prior injection.

2.2 Visual and Text Encoders

Visual Branch. Following TransCP [15], our visual encoder is instantiated as a hybrid architecture comprising a CNN backbone and a vision transformer, synergistically capturing fine-grained local textures and long-range contextual dependencies. Specifically, a ResNet-50 backbone is employed to extract feature maps from the input pathology image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$, denoted as $\mathcal{Z} \in \mathbb{R}^{C_v\times H_v\times W_v}$ where $H_v = H/32$, $W_v = W/32$, $C_v = 256$. The spatial feature map $\mathcal{Z}$ is subsequently flattened and projected into a sequence of visual tokens $\mathcal{Z}_v \in \mathbb{R}^{C_v\times N_v}$ ($N_v = H_vW_v$), and processed by a 6-layer transformer encoder $\mathcal{E}_v$ to derive the visual representation:

$$\mathcal{F}_v = \mathcal{E}_v(\mathcal{Z}_v) \in \mathbb{R}^{C_v\times N_v}. \tag{1}$$

Text Branch. Similar to [18], we use BERT $\mathcal{E}_t$ as the textual backbone. Given a referring expression T_{exp}, the raw text is partitioned into subword tokens and mapped to a high dimensional latent space through the BERT embedding layer. The transformer encoder then computes contextualized textual embeddings, denoted as $\mathcal{F}_w \in \mathbb{R}^{C_w\times N_w}$, where C_w represents the feature dimensionality and N_w is the sequence length. This process is formulated as:

$$\mathcal{F}_w = \mathcal{E}_t(T_{\text{exp}}; \theta_t), \tag{2}$$

where $\mathcal{E}_t(\cdot;\theta_t)$ represents the text encoder's θ_t-parameterized mapping function.

2.3 VWF: Visual-Word Fusion

To leverage visual features $\mathcal{F}_v$ and textual features $\mathcal{F}_w$, we propose the Visual-Word Fusion (VWF) module to refine word level linguistic representations by assimilating image conditioned local visual evidence. This mechanism strengthens fine-grained spatial alignment between pathological descriptors and target regions, and is crucial in localizing inconspicuous targets shrouded by stochastic textures, low contrast, and interference from adjacent morphological structures.

To facilitate multimodal interaction, both visual and textual features are first projected into a unified embedding space with dimension C and reshaped into a token-first form, denoted as $F_v \in \mathbb{R}^{N_v\times C}$ and $F_w \in \mathbb{R}^{N_w\times C}$, respectively. VWF employs text tokens as Query and visual tokens as Key/Value to retrieve word-wise visual evidence:

$$\tilde{F}_f = \text{MHA}(Q = F_w,\ K = F_v,\ V = F_v;\ M_v), \tag{3}$$

where M_v is the visual padding mask. The retrieved visual evidence is subsequently integrated into the textual stream via a gated residual update:

$$F_w^c = \text{LN}\Big(F_w + \sigma(\alpha)\tilde{F}_f\Big), \qquad F_w^c = \text{LN}(F_w^c + \text{FFN}(F_w^c)). \tag{4}$$

Here, α is initialized with a negative bias such that the gating value $\sigma(\alpha)$ remains near zero during the initial training phase. This enforces a conservative fusion strategy that prioritizes linguistic stability early on, while adaptively augmenting the textual features with progressively stronger visual signals as training converges. Invalid text-token outputs are masked out using the text mask M_w. In our coarse to fine design, the VWF module operates within Stage-1 using the original text features $\mathcal{F}_w$ as input. The resulting calibrated features $\mathcal{F}_w^c$ are then passed to Stage-2 of C2F-BAR. This allows the Stage-2 to leverage the visually grounded linguistic cues established in Stage-1 to replace original text features.

2.4 C2F-BAR: Coarse-to-Fine Boundary-Aware Regression

As shown in Fig. 1, **C2F-BAR** is a two-stage regression scheme built on the shared **VLT** regressor. It decouples stable target localization from boundary-aware refinement under box-only supervision. Specifically, Stage-1 establishes a spatial anchor by predicting a coarse bounding box, while Stage-2 iteratively refines this proposal by leveraging visually-enhanced linguistic features (from VWF) synergized with morphological priors derived from SAM3.

VLT: Visual-Language Transformer for Box Regression VLT serves as a shared box regressor shared across both stages of C2F-BAR. It is designed to decode normalized bounding box coordinates by integrating visual features with their corresponding linguistic counterparts.

Let F_f represent the VWF fusion feature. Subsequently, a learnable regression token t_{reg} is prepended to the F_f, which is then processed by a visual language transformer $\mathcal{G}$ to aggregate global contextual information and decode:

$$h = \mathcal{G}(\text{concat}(t_{reg}, F_f); Top_{mask}) \in \mathbb{R}^C, \tag{5}$$

where Top_{mask} is selectively activated to inject spatial priors only during the Stage-2 refinement phase. $h \in \mathbb{R}^C$ denote the output feature of the regression token. The bounding box is predicted as

$$b = \sigma(\text{MLP}(h)) \in [0, 1]^4. \tag{6}$$

This design ensures a parameter-efficient framework where a common VLT architecture is reused, while the specificity of each stage is maintained through distinct textual inputs and spatial biases.

Stage-1: Coarse localization. In Stage-1, the VLT ingests VWF feature F_f, fusing visual features and textual embeddings to regress a coarse box b_0:

$$b_0 = \text{VLT}(\text{concat}(t_{reg}, \mathcal{F}_f); Top_{mask} = 0)\,. \tag{7}$$

This stage prioritizes global localization stability, generating robust spatial proposals as initialization for subsequent fine-grained refinement.

Stage-2: Boundary-aware refinement with CBP and PPT.

Confidence-weighted Boundary Prior(CBP). Stage-2 substitutes the raw text embeddings with VWF-calibrated features $\mathcal{F}_w^c$ to generate F_{f_1} and assimilates an additive boundary prior, B_{mask}, into the attention mechanism of VLT:

$$b_1 = \mathrm{VLT}(\mathrm{concat}(t_{reg}, \mathcal{F}_{f_1}); T_{mask})\,. \tag{8}$$

This boundary prior, distilled from a frozen SAM3 teacher model, provides pixel-level morphological guidance to rectify potential boundary drifts.

To construct this prior, we query SAM3 with a box prompt b_p and retrieve the top-K ($K = 5$) candidate masks along with their confidence scores. These masks are spatially aligned to the visual token grid and projected into the latent space. We then synthesize a unified prior by aggregating these candidates via a confidence-weighted mechanism. To avoid early over-commitment to a potentially noisy candidate, the weighting distribution is scheduled to become progressively sharper during training, so that the prior initially aggregates multiple plausible regions and gradually concentrates on the most reliable one.

Since priors may degrade when prompted with inaccurate boxes, we design a reliability gating mechanism. The final prior injected into Stage-2 is as:

$$T_{mask} = r \cdot \phi_K, \tag{9}$$

where ϕ_K denotes the aggregated token-space prior from the top-K SAM3 candidates B_{mask} and $r \in [0, 1]$ is the reliability weight. This design suppresses noisy priors and stabilizes refinement under weak boundaries and tiny target settings.

Progressive Prompt Transition (PPT). To ensure high quality prior generation throughout the training process, we propose a Progressive Prompt Transition (PPT) strategy for constructing the query box b_p. Since nascent model predictions in early epochs often yield erratic SAM3 candidates, we implement a curriculum learning approach. Specifically, the training phase commences by utilizing the ground-truth boxes as prompts to stabilize the warm up. We then transition to using Stage-1 predictions (b_0) via a linear annealing schedule, where the sampling probability of ground-truth boxes decays over a mixing period. Ultimately, the model relies exclusively on b_0, aligning with the inference protocol.

2.5 Loss Function

We supervise the refined box b_1 and the coarse prediction b_0 in our PCRP. The overall objective is

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\ell_1}(b_1, b_{\mathrm{gt}}) + \lambda_2 \mathcal{L}_{\mathrm{giou}}(b_1, b_{\mathrm{gt}}) + \lambda_3 \mathcal{L}_{\ell_1}(b_0, b_{\mathrm{gt}}) + \lambda_4 \mathcal{L}_{\mathrm{giou}}(b_0, b_{\mathrm{gt}}), \tag{10}$$

where b_{gt} is the ground-truth box, and $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are trade-off weights. $\mathcal{L}_{\ell_1}$ denotes the coordinate regression loss in (x, y, w, h) space, and $\mathcal{L}_{\mathrm{giou}}$ denotes the GIoU loss. In practice, we set $\lambda_1 = \lambda_3 = 5$ and $\lambda_2 = \lambda_4 = 2$.

Table 1: PCRP 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[5]	ICCV'21	RN50/BERT-B	58.40	52.86	68.72	66.75	50.29	41.94
SeqTR[19]	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†[9]	NeuIPS'23	CLIP-L/LLaMa	62.32	57.96	73.52	70.24	53.51	48.31
TransCP[15]	TPAMI'24	RN50/BERT-B	61.73	56.81	74.27	71.92	51.87	44.93
SimVG[4]	NeurIPS'24	ViT-B/BERT-B	63.94	59.42	75.36	73.18	52.52	46.92
D-MDETR [13]	TPAMI'24	CLIP-B/CLIP-B	64.92	57.69	76.29	73.10	55.98	45.57
PKNet [18]	MICCAI'25	RN50/BERT-B	69.95	63.49	80.48	76.88	61.66	52.95
PCRP (Ours)	-	RN50/BERT-B	**81.63**	**72.76**	**85.04**	**80.23**	**78.22**	**66.89**

3 Experiment

Dataset. We evaluate PCRP on RefPath [18], which contains 27,610 histopathology images and 33,500 language-grounded bounding boxes. Following [18], RefPath is split into 24,757 training images with 30,452 boxes and 2,853 test images with 3,048 boxes, each paired with a pathology-related expression. The test set is divided by magnification into testA (40×, 1,342 boxes) and testB (20×, 1,706 boxes): 20× images preserve broader tissue organization and surrounding-cell context, while 40× images reveal finer structural and cellular details.

Evaluation Metric. We follow RefPath [18] and use accuracy (Acc%) as the primary metric [18]. Since pathological images at different magnifications show clear differences in target scale and boundary appearance, applying a single IoU threshold across magnifications may lead to biased comparisons. Therefore, we adopt the RefPath settings by using an IoU threshold of 0.7 for 40× images and 0.5 for 20× images. In addition, to complement the single-threshold metric, we also report mIoU% to provide a more complete assessment of grounding quality.

Implementation details. We train and test our model on two NVIDIA GeForce RTX 4090 GPUs. The CNN backbone (ResNet-50) and visual Transformer encoder are initialized from pre-trained DETR-R50 weights. We use AdamW with an initial learning rate of 1×10^{-5} and decay the learning rate at epoch 90. The model is trained for 120 epochs. The batch size is 32. To improve grounding across magnifications, especially on 20× images, we use a progressive visual cue scheme. During training, we warm up with GT prompts for 5 epochs and then mix GT and predicted prompts over the next 20 epochs.

Results on RefPath. Tab. 1 compares Acc and mIoU on RefPath with representative baselines, including TransVG [5], SeqTR [19], CLIPVG [17], LLaVa-Med† [9], TransCP [15], SimVG [4], D-MDETR [13], and PKNet [18]. PCRP achieves the best results on all splits: 81.63/72.76 on $RefPath_{all}$, 85.04/80.23 on $RefPath_{40\times}$, and 78.22/66.89 on $RefPath_{20\times}$ (Acc/mIoU), with the largest margin on $RefPath_{20\times}$. Compared with PKNet [18], PCRP improves by 11.68/9.27 on $RefPath_{all}$ and 16.56/13.94 on $RefPath_{20\times}$. This gain mainly comes from mask priors: a coarse prediction provides candidate regions for SAM3 masks, and refinement uses these masks to guide boundary regression, reducing

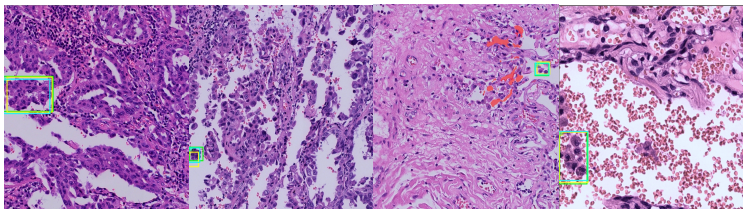

Fig. 2: Qualitative visualization of **PCRP**. Green denotes the ground-truth box, yellow denotes the Stage-1 prediction, and blue denotes the Stage-2 prediction.

Table 2: Ablation of **PCRP**. All variants use the same Visual Branch, Text Branch, and VLT regressor, and progressively enable **VWF**, **CBP**, and **PPT**.

ID	VWF	CBP	PPT	$RefPath_{all}$ Acc/mIoU ↑	$RefPath_{40\times}$ Acc/mIoU ↑	$RefPath_{20\times}$ Acc/mIoU ↑
(a)				78.90/69.40	82.40/78.20	75.30/63.40
(b)	✓			79.85/70.20	82.95/78.55	76.90/65.10
(c)	✓	✓		81.30/72.45	84.55/80.10	78.15/66.70
(d)	✓	✓	✓	**81.63/72.76**	**85.04/80.23**	**78.22/66.89**

spatial drift and yielding tighter localization at low magnification. $RefPath_{20\times}$ remains hardest because cues are weaker and boundaries less distinct.

Qualitative Results. As shown in Fig. 2, the Stage-1 box (yellow) is offset from the ground-truth box (green), reflecting the difficulty of stabilizing word-to-region correspondence in pathology images with textures. With Stage-2 refinement (blue), leveraging visually grounded language cues and boundary priors, the prediction is pulled back toward the target region, demonstrating reduced grounding drift and improved localization in weak-boundary cases.

Ablation Study. Tab. 2 reports **PCRP** ablations on RefPath. All variants share the same Visual Branch, Text Branch, and VLT regressor, differing only in **VWF** and Stage-2 controls of **C2F-BAR**. Enabling **VWF** ((a)→(b)) yields consistent gains, especially on $RefPath_{20\times}$, indicating better word-to-region alignment for low-magnification small targets. Adding **CBP** ((b)→(c)) further improves performance, especially mIoU, showing that reliability-aware prior weighting stabilizes boundary refinement under box-only supervision.Adding **PPT** ((c)→(d)) performs best, suggesting that progressive prompt transition improves prior quality during training and benefits Stage-2 refinement.

Failure cases and computational overhead. PCRP corrects moderate Stage-1 errors via boundary-aware refinement, but poor boxes or difficult lesions can misguide SAM3 and yield unreliable priors, causing distractors and unstable masks in Fig. 2. SAM3 dominates overhead, raising runtime/memory from 40.2 ms/959.9 MB to 153.7 ms/5771.9 MB, with 840.5M parameters.

4 Conclusion

We propose PCRP for pathology visual grounding. The framework consists of two core modules. **VWF** targets semantic drift by injecting word-wise visual evidence into word-level text features, strengthening word-to-region alignment in low-magnification small-target cases. **C2F-BAR** addresses boundary ambiguity under box-only supervision via coarse-to-fine refinement with reliability-controlled mask priors. Experimental results on RefPath verify effectiveness of the framework and show larger gains in low-magnification small-target scenarios.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grants 2023YFC2705704, 2023YFC2705705, the Fundamental Research Funds for the Central Universities (No. 2042026kf0044), the National Natural Science Foundation of China (Grant No. 62471342), The Innovative Research Group Project of Hubei Province under Grants 2024AFA017, the New Cornerstone Science Foundation through the XPLORER PRIZE and WHU-Kingsoft Joint Lab.

Disclosure of Interests

We have no conflicts of interest to disclose.

References

1. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Coll-Vinent, D.S., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., HAZRA, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollar, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: SAM 3: Segment anything with concepts. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=r35clVtGzw
3. Chen, Z., Zhou, Y., Tran, A., Zhao, J., Wan, L., Ooi, G.S.K., Cheng, L.T.E., Thng, C.H., Xu, X., Liu, Y., et al.: Medical phrase grounding with region-phrase context contrastive alignment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 371–381. Springer (2023)
4. Dai, M., Yang, L., Xu, Y., Feng, Z., Yang, W.: Simvg: A simple framework for visual grounding with decoupled multi-modal fusion. Advances in Neural Information Processing Systems **37**, 121670–121698 (2025)

5. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: Transvg: End-to-end visual grounding with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1769–1779 (2021)
6. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
7. Huang, X., Huang, H., Shen, L., Yang, Y., Shang, F., Liu, J., Liu, J.: A refer-and-ground multimodal large language model for biomedicine. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 399–409. Springer (2024)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
9. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
10. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
11. Matasyoh, N.M., Schmidt, R., Zeineldin, R.A., Spetzger, U., Mathis-Ullrich, F.: Interactive surgical training in neuroendoscopy: Real-time anatomical feature localization using natural language expressions. IEEE Transactions on Biomedical Engineering **71**(10), 2991–2999 (2024)
12. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
13. Shi, F., Gao, R., Huang, W., Wang, L.: Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(2), 1181–1198 (2023)
14. Sun, Y., Wu, H., Zhu, C., Zheng, S., Chen, Q., Zhang, K., Zhang, Y., Wan, D., Lan, X., Zheng, M., et al.: Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In: European Conference on Computer Vision. pp. 56–73. Springer (2024)
15. Tang, W., Li, L., Liu, X., Jin, L., Tang, J., Li, Z.: Context disentangling and prototype inheriting for robust visual grounding. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(5), 3213–3229 (2023)
16. Wang, H., Wu, H., Qin, J.: 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. pp. 11408–11417 (2024)
17. Xiao, L., Yang, X., Peng, F., Yan, M., Wang, Y., Xu, C.: Clip-vg: Self-paced curriculum adapting of clip for visual grounding. IEEE Transactions on Multimedia **26**, 4334–4347 (2023)
18. Zhong, C., Hao, S., Wu, J., Chang, X., Jiang, J., Nie, X., Tang, H., Bai, X.: Pathvg: A new benchmark and dataset for pathology visual grounding. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 454–463. Springer (2025)
19. Zhu, C., Zhou, Y., Shen, Y., Luo, G., Pan, X., Lin, M., Chen, C., Cao, L., Sun, X., Ji, R.: Seqtr: A simple yet universal network for visual grounding. In: European Conference on Computer Vision. pp. 598–615. Springer (2022)