Prompt Group-Aware Training for Robust Text-Guided Nuclei Segmentation

Yonghuang Wu[1*], Zhenyang Liang[1*], Wenwen Zeng[1], Xuan Xie[1], Jinhua Yu[1]

Fudan University, Shanghai, China
jhyu@fudan.edu.cn
* These authors contributed equally to this work.

Abstract. Foundation models such as Segment Anything Model 3 (SAM3) enable flexible text-guided medical image segmentation, yet their predictions remain highly sensitive to prompt formulation. Even semantically equivalent descriptions can yield inconsistent masks, limiting reliability in clinical and pathology workflows.

We reformulate prompt sensitivity as a group-wise consistency problem. Semantically related prompts are organized into *prompt groups* sharing the same ground-truth mask, and a prompt group-aware training framework is introduced for robust text-guided nuclei segmentation. The approach combines (i) a quality-guided group regularization that leverages segmentation loss as an implicit ranking signal, and (ii) a logit-level consistency constraint with a stop-gradient strategy to align predictions within each group. The method requires no architectural modification and leaves inference unchanged.

Extensive experiments on multi-dataset nuclei benchmarks show consistent gains under textual prompting and markedly reduced performance variance across prompt quality levels. On six zero-shot cross-dataset tasks, our method improves Dice by an average of 2.16 points. These results demonstrate improved robustness and generalization for vision-language segmentation in computational pathology.

Keywords: Foundation Models · Prompt Robustness · Preference-Aware Learning · Text-Guided Segmentation.

1 Introduction

Foundation models have reshaped image segmentation by introducing promptable and highly generalizable architectures. Segment Anything (SAM) [7] enables a single model to support diverse prompts, achieving strong zero-shot performance across domains. This flexibility has motivated growing interest in applying SAM-like models to medical and pathology imaging.

However, recent studies reveal that prompt-driven segmentation models are highly sensitive to prompt formulation, particularly in medical settings [16]. Semantically equivalent text prompts may produce inconsistent segmentation outcomes. In pathology images, prompts such as “nuclei”, “all cell nuclei”, or

implicit subtype descriptions often refer to the same target but lead to unstable predictions, undermining reliability for clinical deployment.

Text-guided and open-vocabulary segmentation methods further explore language-conditioned dense prediction. CLIP-based approaches [13, 9, 3] and medical vision-language models [21] align visual and textual representations to enable segmentation from free-form descriptions. Despite their success, these methods typically assume a one-to-one correspondence between a prompt and its target region, and performance remains sensitive to prompt phrasing. Prompt variability is rarely modeled explicitly during supervised training.

Related robustness studies address annotation uncertainty, noisy user interactions, or adversarial prompts [23, 10]. Yet ambiguity is generally treated as noise to be mitigated, rather than as structured equivalence among multiple valid prompts describing the same segmentation target. In pathology, linguistic variability is intrinsic: different expressions may legitimately correspond to identical anatomical or cellular structures, naturally inducing a many-to-one mapping from prompts to masks.

To address this gap, we reformulate prompt sensitivity as a group-wise consistency problem (see 1). For each image, we organize semantically related prompts into *prompt groups* that share a common ground-truth mask. We propose a group-aware training framework that enforces prompt-invariant behavior through two mechanisms: (i) a quality-guided objective that models relative prompt reliability within each group, and (ii) a prompt consistency loss that aligns segmentation outcomes across prompts. Our method operates purely at training time, requires no additional supervision, and leaves the inference procedure unchanged.

Experiments on text-guided nuclei segmentation demonstrate that our approach improves accuracy while substantially reducing performance variance across diverse prompts, providing a practical pathway toward robust and trustworthy vision-language models in pathology.

2 Method

2.1 Task Definition and Prompt Grouping

We study referring image segmentation in pathology images under varying prompt quality. Given an image $I \in \mathbb{R}^{H\times W\times C}$ and a set of natural language prompts $\mathcal{P} = \{p_1, \ldots, p_K\}$ that refer to the same target structure, the goal is to predict a prompt-invariant binary segmentation mask. A text-conditioned segmentation model produces one mask per prompt:

$$S_i = f_\theta(I, p_i), \quad S_i \in \{0,1\}^{H\times W}. \tag{1}$$

In pathology, semantically equivalent prompts may differ substantially in clarity or specificity (e.g., "nuclei" vs. implicit subtype descriptions), making text a potentially noisy conditioning signal. To explicitly model prompt equivalence, we organize training data into *prompt groups*. Each group is defined as

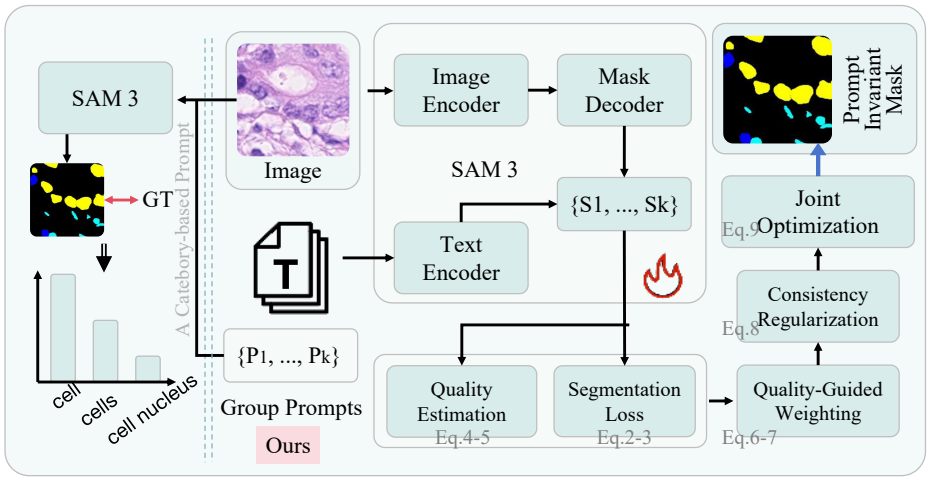

Fig. 1. Overview of the proposed prompt group-aware training framework.

$(I, \mathcal{P}_g, M_g)$, where all prompts $\mathcal{P}_g$ correspond to the same ground-truth mask M_g, inducing a many-to-one mapping from prompts to supervision.

Prompt groups naturally include (i) **all-instance groups**, targeting all structures of interest, and (ii) **class-specific groups**, focusing on a particular cell type. This formulation enables prompt-invariant learning by sharing supervision across diverse linguistic expressions, while mitigating conflicting gradients from semantically equivalent prompts.

2.2 Group-wise Segmentation and Prompt Quality Modeling

For a given image I and prompt group $\mathcal{P}_g$, the model produces one prediction per prompt:

$$S_i = f_\theta(I, p_i). \tag{2}$$

Each prediction is optimized with the standard SAM-style segmentation loss:

$$\mathcal{L}_{seg}^{(i)} = \mathcal{L}_{mask}^{(i)} + \mathcal{L}_{dice}^{(i)} + \mathcal{L}_{presence}^{(i)}, \tag{3}$$

The image encoder is shared across prompts, while the text encoder and mask decoder are applied independently, preserving the original architecture and training protocol.

Prompt Quality Estimation Prompts in the same group may vary in specificity. We quantify prompt quality using the segmentation loss:

$$q_i = -\mathcal{L}_{seg}^{(i)}, \tag{4}$$

where $\mathcal{L}_{seg}^{(i)}$ is computed using the same top-K selection and max-aggregation strategy as in inference.

We consider only relative quality within each prompt group:

$$\tilde{q}_i = q_i - \frac{1}{K}\sum_{j=1}^{K} q_j. \tag{5}$$

The quality scores are detached from the computational graph and used only as a fixed relative ranking signal within each group.

Quality-Guided Group-aware Loss We further define a soft weighting scheme to modulate the contribution of different prompts:

$$w_i = \frac{\exp(-\mathcal{L}_{seg}^{(i)}/\tau)}{\sum_{j=1}^{K}\exp(-\mathcal{L}_{seg}^{(j)}/\tau)}, \tag{6}$$

where τ is a temperature hyperparameter.

Using the relative prompt quality $\tilde{q}_i$, we define a group-aware regularization objective:

$$\mathcal{L}_{group} = -\sum_{i=1}^{K} \tilde{q}_i \log w_i. \tag{7}$$

This regularization aligns the learned weights with detached relative prompt quality, while preventing the ranking signal itself from being directly optimized.

2.3 Prompt Consistency Regularization and Training Objective

While quality-aware weighting adjusts prompt importance, it does not enforce prediction consistency. We therefore introduce a logit-level consistency regularization.

Let Z_i denote the predicted mask logits (before sigmoid) for prompt p_i. We select the first prompt in the group as a reference and define:

$$\mathcal{L}_{cons} = \frac{1}{K-1}\sum_{i=2}^{K} \left\| Z_i - \mathrm{stopgrad}(Z_1) \right\|_2^2. \tag{8}$$

This objective encourages consistent predictions across prompts. A stop-gradient is applied only to the reference logit to avoid mutual reinforcement, while gradients are preserved for the remaining prompts. Prompts are shuffled during training to prevent bias from a fixed reference order.

Overall Training Objective and Inference The final training objective for a prompt group is:

$$\mathcal{L} = \frac{1}{K}\sum_{i=1}^{K} \mathcal{L}_{seg}^{(i)} \; + \; \lambda \mathcal{L}_{group} \; + \; \beta \mathcal{L}_{cons}, \tag{9}$$

where λ and β control the strength of quality-aware weighting and prompt consistency enforcement, respectively.

3 Experiments

3.1 Datasets and Referring Protocol

All experiments follow the segmentation protocol of [22]. Models are trained on PanNuke [4] and CoNSeP [5] using only 10% of training images to simulate data-efficient clinical scenarios. Evaluation is conducted on unseen datasets (CPM15, CPM17 [20], Histology [6], Kumar [8], CryoNuSeg [15]) under strict cross-dataset generalization.

Each image–target pair is associated with a prompt group, consisting of multiple textual prompts that refer to the same segmentation target and therefore share an identical ground-truth mask. We consider two task settings throughout the experiments: T1 (all-nuclei segmentation), where prompts refer to all nuclei in the image, and T2 (category-specific segmentation), where prompts refer to nuclei of a specific cell category. Prompt groups are generated in a controlled and reproducible manner by varying linguistic completeness and specificity while keeping supervision unchanged.

Prompts are constructed along two orthogonal dimensions: prompt type (T1: all nuclei vs. T2: category-specific nuclei) and prompt quality, which controls the level of linguistic detail. We define three prompt quality levels: low, medium, and high (see Table 3). Lower-quality prompts are intentionally short and underspecified, whereas higher-quality prompts provide richer semantic and contextual constraints. This design enables controlled and reproducible analysis of robustness to prompt variation without introducing artificial noise.

3.2 Experimental Setup and Baselines

All models use AdamW ($\mathrm{lr} = 1 \times 10^{-4}$, batch size 1, 5 epochs) on identical data splits. During training, prompt groups with K prompts are randomly sampled per image. All prompts within the same group correspond to alternative linguistic descriptions of the *same* segmentation target and share an identical ground-truth mask; thus, this setting does not introduce additional labeled data or extra semantic supervision. In this sense, multiple prompts act as structured language-space augmentations that enforce prompt-invariant predictions, increasing the number of consistency constraints rather than the amount of supervision. All methods are evaluated using a single prompt at inference time without ensembling. Default hyperparameters: $\tau = 1.0$, $\lambda = 1.0$, $\beta = 0.1$. All experiments run on a single MetaX C500 GPU (64 GB), ensuring practical reproducibility.

Baselines. **Vision-prompt methods**: HSAM [2], SAN [19], InstaSAM [17], MedSAM [14] (trained separately for T1/T2; results from prior work). **Text-prompt SOTA**: CLIP-Seg [13], Grounded-SAM2 [18], SAM3 [1] (fully fine-tuned on identical schedules; reported separately for T1/T2 for fair comparison). **Large autoregressive models**: SegZero [11], VisionReasoner [12] (test-only evaluation due to prohibitive computational costs).

3.3 Quantitative Results

Table 1 reports quantitative results under visual and textual prompting. Under text prompts, our method performs best on both PanNuke (79.42 / 62.01) and CoNSeP (76.81 / 46.86), surpassing the strongest text baseline SAM3* by +0.97/+6.20 and +1.78/+3.24 Dice (T1/T2), respectively. Improvements are more pronounced for category-specific segmentation (T2), reflecting stronger fine-grained semantic grounding. Compared with prior text-driven and autoregressive models, our approach consistently achieves higher accuracy while relying solely on textual prompts at inference. Qualitative examples are shown in Figure 2.

Table 1. Segmentation performance (Dice ↑) under different methods. Visual and Text columns indicate the type of prompts used for inference. For text-prompt methods, results are reported in the format "T1 / T2", where T1 denotes all-nuclei segmentation and T2 denotes category-specific nuclei segmentation.

Method	PanNuke		CoNSeP	
	Visual	Text	Visual	Text
H-SAM	64.90 / 40.36	-	18.56 / 29.76	-
SAN	67.77 / 40.41	-	-	-
InstaSAM	74.67 / 21.21	-	-	-
MedSAM	78.43 / 47.11	-	21.23 / 31.35	-
CLIP-Seg*	-	65.53 / 47.19	-	63.18 / 37.90
GroundedSAM2*	-	73.60 / 55.81	-	72.35 / 43.62
Seg-Zero	-	26.01 / 12.75	-	22.04 / 10.57
VisionReasoner	-	47.95 / 17.90	-	45.18 / 17.77
SAM 3	-	76.53 / 42.63	-	68.15 / 26.22
SAM 3*	-	78.45 / 48.98	-	75.03 / 40.89
Ours	-	**79.42 / 62.01** +0.97 / +6.2	-	**76.81 / 46.86** +1.78 / +3.24

3.4 Ablation and Sensitivity Analysis

We conduct ablation studies to validate key design choices and analyze model sensitivity to hyperparameters.

Group-aware supervision is essential. Removing $\mathcal{L}_{group}$ and $\mathcal{L}_{cons}$ drops performance to 78.45 / 48.98, demonstrating that per-prompt supervision alone cannot handle prompt variability. The consistency loss further improves alignment, with performance degrading to 77.72 / 54.30 when removed. Full pairwise consistency (77.38 / 51.97) underperforms the proposed stop-gradient formulation, suggesting that naive all-to-all alignment introduces optimization conflicts. Our stability-aware design mitigates this by using a single reference, achieving better convergence. Performance peaks around $K = 3$-4 and gradually declines for $K \geq 5$, indicating diminishing returns from additional prompt variations.

Table 2. Ablation and sensitivity analysis on low-quality prompts (Dice ↑). Results are reported in the format "T1 / T2" (all-nuclei / category-specific).

Loss design		*Hyperparameter sensitivity*	
Setting	Dice	Setting	Dice
Full model	**79.42 / 62.01**		
w/o $\mathcal{L}_{group}, \mathcal{L}_{cons}$	78.45 / 48.98	$K = 2$	78.10 / 61.52
w/o $\mathcal{L}_{cons}$	77.72 / 54.30	$K = 3$	**79.42 / 62.01**
$\mathcal{L}_{cons}$ (full pairwise)	77.38 / 51.97	$K = 4$	79.92 / 61.24
$\beta = 0.05$	78.50 / 61.36	$K = 5$	79.11 / 60.75
$\beta = 0.2$	78.06 / 60.86	$K = 6$	78.41 / 60.14

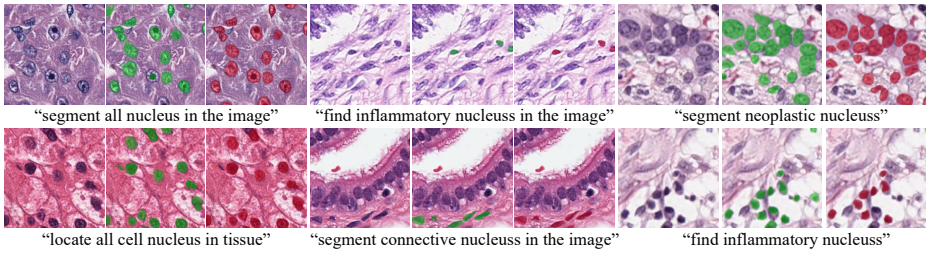

Fig. 2. Qualitative comparison of segmentation results under different prompts.

The method remains stable across this range, with the loss weight β showing robustness around the default value of 0.1.

3.5 Robustness to Prompt Quality Variation

To analyze robustness under linguistic variation, we evaluate all methods across structured prompt quality levels (`low`, `medium`, `high`). As shown in Table 4, baseline methods exhibit pronounced performance degradation as prompt quality decreases. In contrast, our method degrades more gracefully and consistently maintains higher segmentation accuracy, with the largest gains observed under low-quality prompts.

Zero-shot cross-domain generalization. We further evaluate zero-shot generalization on multiple external datasets covering diverse tissue types, imaging modalities, and acquisition protocols. None of the evaluated models are trained or fine-tuned on these target datasets.

SAM2/SAMPO use standard visual prompts, while SAM3 and ours rely solely on text with identical construction across datasets. As shown in Table 5, our method achieves the best results on most benchmarks (Histology, CPM15/17, Kumar), remaining competitive with vision-prompt-based SAMPO. Performance slightly decreases on CryoNuSeg, yet overall results demonstrate strong zero-shot transferability.

Table 3. Overview of prompt types and quality levels used in our referring segmentation protocol. All prompts within the same group refer to the same segmentation target and share identical ground-truth masks.

Dimension	Category	Description
Prompt type	All	Refers to all nuclei in the image without category distinction (e.g., "all nuclei", "all cell nuclei in the tissue")
	Category	Refers to nuclei of a specific cell category (e.g., inflammatory nuclei, epithelial cell nuclei)
Prompt quality	Low	Short and underspecified prompts with limited linguistic cues; typically contain a single verb and object
	Medium	Moderately descriptive prompts with paraphrasing or partial specification, including optional spatial context
	High	Explicit and detailed prompts with richer verbs, spatial constraints, and occasional exclusion clauses

Table 4. Segmentation performance under different prompt quality levels. Low, Medium, and High denote prompts with increasing linguistic specificity and semantic completeness (see Table 3). Results are reported in the format "T1 / T2" (all-nuclei / category-specific).

Methods	Low	Medium	High
SAM3	77.28 / 47.20	78.75 / 47.64	78.64 / 48.23
Ours	**78.75** / **62.54**	**78.77** / **62.80**	**78.89** / **61.86**

Table 5. Zero-shot cross-domain generalization results for **all-nuclei segmentation**. Best results are shown in **bold**. Kumar-Same and Kumar-Diff denote test sets with organ types seen and unseen during training, respectively.

Methods	Histology	CPM15	CPM17	Kumar-same	Kumar-diff	CryoNuSeg
SAM2	46.91	43.10	42.29	36.29	40.77	35.78
SAM3	52.46	56.07	57.87	34.53	53.22	64.10
SAMPO	70.16	79.28	81.14	78.45	83.19	77.94
Ours	**74.61** (+4.45)	**79.56** (+0.28)	**86.10** (+4.96)	**80.50** (+2.05)	**84.78** (+1.59)	**77.58** (-0.36)

4 Conclusion and Limitations

We present a prompt group–aware training framework that improves robustness in text-guided medical image segmentation by modeling semantically equivalent prompts with a shared target. Experiments on nuclei segmentation show enhanced stability across diverse textual descriptions, indicating good generalizability.

Our approach adopts a fixed text encoder to enable controlled and efficient analysis of prompt variability, which may limit the modeling of highly complex semantics. Future work will explore integrating more expressive text encoders, such as large language models, and developing more advanced preference-based optimization strategies to further enhance robustness and semantic understanding.

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
2. Cheng, Z., Wei, Q., Zhu, H., Wang, Y., Qu, L., Shao, W., Zhou, Y.: Unleashing the potential of sam for medical adaptation via hierarchical decoding. In: CVPR (2024)
3. Ding, Z., Wang, J., Tu, Z.: Open-vocabulary universal image segmentation with maskclip. arXiv preprint arXiv:2208.08984 (2022)
4. Gamper, J., Koohbanani, N.A., Benes, K., Graham, S., Jahanifar, M., Khurram, S.A., Azam, A., Hewitt, K., Rajpoot, N.: Pannuke dataset extension, insights and baselines. arXiv preprint arXiv:2003.10778 (2020)
5. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.: Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical Image Analysis p. 101563 (2019)
6. He, Z., Unberath, M., Ke, J., Shen, Y.: Transnuseg: A lightweight multi-task transformer for nuclei segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 206–215. Springer (2023)
7. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
8. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE transactions on medical imaging **36**(7), 1550–1560 (2017)
9. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546 (2022)
10. Li, H., Liu, H., Hu, D., Wang, J., Oguz, I.: Prism: A promptable and robust interactive segmentation model with visual prompts. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 389–399. Springer (2024)
11. Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025)
12. Liu, Y., Qu, T., Zhong, Z., Peng, B., Liu, S., Yu, B., Jia, J.: Visionreasoner: Unified visual perception and reasoning via reinforcement learning. arXiv e-prints pp. arXiv–2505 (2025)

13. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7086–7096 (2022)
14. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
15. Mahbod, A., Schaefer, G., Bancher, B., Löw, C., Dorffner, G., Ecker, R., Ellinger, I.: Cryonuseg: A dataset for nuclei instance segmentation of cryosectioned h&e-stained histological images. Computers in biology and medicine **132**, 104349 (2021)
16. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. Medical Image Analysis **89**, 102918 (2023)
17. Nam, S., Namgung, H., Jeong, J., Luna, M., Kim, S., Chikontwe, P., Park, S.H.: Instasam: Instance-aware segment any nuclei model with point annotations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 232–242. Springer (2024)
18. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
19. Sun, X., Liu, J., Ren, Y., Xu, X., Shen, Y.: Segment any nuclei: A prompt-free segment anything model for nuclei from histology image. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3726–3730. IEEE (2024)
20. Vu, Q.D., Graham, S., Kurc, T., To, M.N.N., Shaban, M., Qaiser, T., Koohbanani, N.A., Khurram, S.A., Kalpathy-Cramer, J., Zhao, T., et al.: Methods for segmentation and classification of digital microscopy tissue images. Frontiers in bioengineering and biotechnology **7**, 53 (2019)
21. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887 (2022)
22. Wu, Y., Zeng, W., Xie, X., Zhao, C., Wu, G., Yu, J.: Sampo: Visual preference optimization for intent-aware segmentation with vision foundation models. arXiv preprint arXiv:2508.02464 (2025)
23. Zhang, L., Tanno, R., Xu, M., Huang, Y., Bronik, K., Jin, C., Jacob, J., Zheng, Y., Shao, L., Ciccarelli, O., et al.: Learning from multiple annotators for medical image segmentation. Pattern Recognition **138**, 109400 (2023)