Exploring Efficient Weakly Supervised Ultrasound Video Object Segmentation with VLM-Guided Structure Advantage Learning

Jialu Li[1], Zekai Liu[1], Hongqiu Wang[1], Huihui Xu[1], Qiong Wang[3], and Lei Zhu[1,2]✉

[1] The Hong Kong University of Science and Technology (Guangzhou), China
[2] The Hong Kong University of Science and Technology, Hong Kong, China
[3] Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

Weakly supervised methods aim to reduce the burden of dense manual labeled annotations by using coarse supervision, yet they still rely on intensive human-provided labels, which remain particularly labor-intensive given the large number of blurred ultrasound frames. Moreover, recent coarse supervisions may further lead to performance degradation, especially for blurred lesion boundaries and motion artifacts. Motivated by the strong semantic understanding of recent vision–language models (VLMs) in medical imaging, we propose a novel weakly supervised method that effectively regards the VLM-embedded prior knowledge as the high-level supervisor to guide the segmentation model in autonomously exploring accurate tumor masks. Specifically, we propose a Guidance Conditional Diffusion Strategy to adaptively integrate multiple lesion cues for learning lesion characteristics. Then, we propose an efficient Structure Advantage Learning strategy that encourages the model to actively explore optimal boundary details and overcome the limitations of weak supervision. We further collect a large-scale dataset with 433 annotated breast lesion videos to facilitate research on this topic. Extensive experiments on 3 video datasets demonstrate that our method consistently outperforms recent state-of-the-art methods while substantially reducing the manual annotation workload(link).

Keywords: Weakly supervised video object segmentation · Ultrasound video · Relative advantage-guided learning· Vision-language model

1 Introduction

Automatic segmentation of breast lesions in ultrasound videos is a challenging task, due to blurred lesion boundaries, motion artifacts, and obvious temporal variations across frames [9–11, 19]. Recent segmentation methods mainly rely on accurate pixel-level annotations to ensure robustness and consistency across frames, imposing a labor-intensive and expertise-dependent burden.

Weakly supervised video object segmentation (WSVOS) [23] aims to reduce annotation workload by replacing dense pixel-level annotation with limited

✉ Lei Zhu (leizhu@hkust-gz.edu.cn) is the corresponding author of this work.

weak supervision. Existing WSVOS methods [14, 26] typically leverage intensive per-frame coarse annotations to train the model, whereas box-supervised methods [6,13] rely primarily on manually defined bounding boxes and low-level visual constraints. Hence, these methods could be annotation-intensive for the large number of ultrasound frames. Moreover, solely leveraging these simple visual constraints could be insufficient to disambiguate blurred lesion boundaries and further struggle to recover precise lesion structures.

Recent VLMs and their reinforcement learning (RL) methods have shown promising performance in various token-level optimization tasks. They propose to quantify the relative advantages among different candidate tokens to enhance the VLM high-level reasoning and decision exploration performance [8,21]. However, recent RL methods are not well-suited for pixel-level segmentation under weak supervision, limiting their effectiveness in refining fine-grained lesion predictions in ultrasound videos.

Motivated by the strong semantic understanding of recent VLMs in medical imaging [5, 12, 18], we first revisit the weakly supervised learning paradigm and collect a large-scale ultrasound VOS dataset with 433 annotated breast lesion videos to facilitate research on this topic. We then propose a novel weakly supervised method that leverages expert knowledge from a frozen VLM [22] with minimal human correction as supervision to substantially reduce manual annotation workload. Furthermore, inspired by the core thinking that quantifying the relative advantages among candidate predictions [8, 21], rather than a strict and complex RL pipeline. We propose the Structure Advantage Learning that enables the segmentation model to autonomously assess the reliability of multiple guidance cues and thus progressively explore optimal lesion structure.

The main contributions of this work can be summarized as follows:

- We propose a VLM-driven weakly supervised method that leverages expert knowledge from a frozen VLM to alleviate the manual annotation workload.
- We introduce the Structure Advantage Learning, in which an expert VLM progressively guides the model to learn optimal lesion structures.
- We propose a Guidance Conditional Dirichlet Diffusion Strategy (GCDS) to alleviate boundary ambiguity caused by weak supervision.
- We collect a large-scale annotated ultrasound VOS dataset (433 videos) and experiments on 3 video datasets show that our method consistently outperforms recent state-of-the-art methods.

2 Method

As shown in Fig. 1, to alleviate the dense manual-labeled workload of recent methods [6, 14, 26], we proposed to first regard the VLM-suggested [22] lesion BBoxes and their confidence scores as weak supervision, where manual correction is only required for a small quantity of unconfident BBoxes to save the human workload (Table 5). During training, a normal diffusion model [25] is used to leverage multiple visual cues together with VLM-embedded prior knowledge [22, 24] to generate predictions. Then, to learn refined lesion structure caused by weak

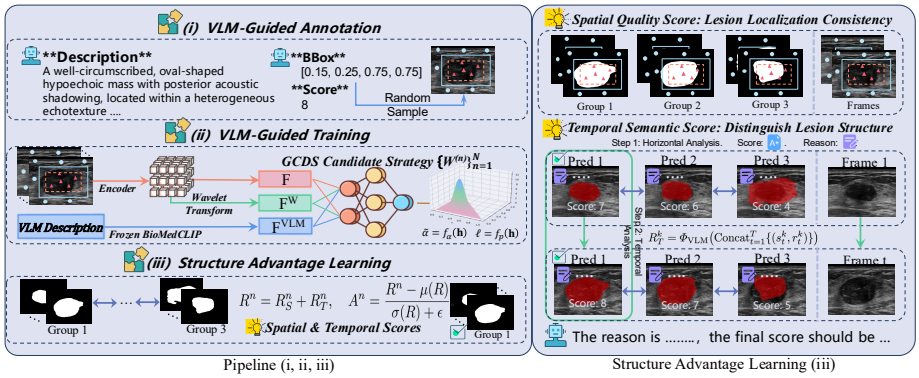

Fig. 1. Overview of our proposed method that leverages the VLM-guided weak supervision with the Structure Advantage learning to alleviate the annotation workload and progressively explore a refined prediction.

supervision, we further propose the Structure Advantage Learning strategy along with the GCDS model that enables the segmentation model to autonomously assess the reliability of multiple guidance cues and progressively explore refined lesion prediction at each denoising timestep.

2.1 VLM-Guided Weak Annotation

In the WSVOS setting, pixel-wise supervision is unavailable. Here, we shrink and enlarge the VLM-suggested coarse lesion BBox (Y_{VLM}) [22], and randomly sample reliable foreground-background point sets $\mathcal{P}$ to construct a coarse pseudo point mask Y_c. We then use the normal diffusion segmentation model [25] as the backbone. In the weakly supervised task, we use the point mask Y_c and the proposed GCDS to assist model training. All unsampled pixels are ignored in the loss computation:

$$\mathcal{P} = \{(x_c, y_c)\}_{c=1}^{C} \sim \text{Sample}(Y_{VLM}, \lambda), \quad Y_c(u, v) = (\mathbf{0}, \mathbf{1}, \mathbf{2})[(u, v) \in \mathcal{P}]. \quad (1)$$

where λ is a hyperparameter to control the BBox sample scale of the background and foreground from Y_{VLM}. Numbers 0, 1, and 2 denote the background, foreground, and unsampled pixels, respectively.

Given each timestep s, to further mitigate the limited coarse supervision Y_c, GCDS strategy $\mathbf{W}^{(n)}$ is proposed to help the diffusion model adaptively leverage multiple guidance cues, including spatial-domain features F, their high-frequency [1,2] wavelet-domain features F^W, and high-level VLM semantic knowledge F^{VLM}:

$$Y_{pred} = \Phi(\mathbf{W}^{(n)}, F, F^W, F^{VLM}, s) \quad (2)$$

Although coarse supervision Y_c can provide limited foreground–background cues to distinguish the lesion object, the boundary structures around the BBox

boundaries remain ambiguous, further limiting the accuracy of lesion supervision in recent weakly supervised methods, especially in ultrasound videos with blurred boundaries and motion artifacts. Furthermore, the reliability of different guidance cues varies across frames under the weak supervision scenario.

Hence, we propose the GCDS with Structure Advantage Learning, which regards the VLM as a high-level supervisor to progressively guide the segmentation model in effectively exploring an optimal tumor structure.

2.2 VLM-Guided Diffusion Training

To achieve precise lesion prediction under weak supervision, the Guidance Conditional Dirichlet Diffusion Strategy (GCDS) is proposed to predict the reliability weights $\mathbf{W}$ for guidance cues, which are then adaptively integrated for the denoising decoder. Specifically, GCDS regards each denoising timestep as a state and, conditioned on this state, samples N groups of candidate reliability combinations $\mathbf{W}$ as the strategies for these guidance cues. This operation is naturally suitable for non-negative, sum-to-one weights [4, 15], further offering an interpretable and flexible strategy for adaptive guidance integration.

Given G guidance cues $\{F^g\}_{g=1}^{G}$, each of shape (B, C, H, W), the mode embedding $\mathbf{h}$ is first derived to initialize a basic Mixed Dirichlet distribution:

$$\mathbf{h} = f_{conv}\Big(\text{Concat}\left(\mathbf{F}_t^{(1)}, \ldots, \mathbf{F}_t^{(G)}\right)\Big). \tag{3}$$

Hence, the initial basic Mixed Dirichlet parameters, mode confidence $\mathbf{p}_0$ and basic concentrations $\boldsymbol{\alpha}_0$ [15], can be formulated as :

$$\begin{aligned} \ell &= f_p(\mathbf{h}) \in \mathbb{R}^{B\times K}, \qquad \mathbf{p}_0 = \text{softmax}(\ell), \\ \tilde{\alpha} &= f_\alpha(\mathbf{h}) \in \mathbb{R}^{B\times (KG)}, \quad \boldsymbol{\alpha}_0 = \text{Softplus}\big(\text{reshape}(\tilde{\alpha},\, B,\, K,\, G)\big). \end{aligned} \tag{4}$$

where f_p and f_α are implemented by global average pooling followed by MLP.

To ensure the GCDS better approximates and learns the reliability of each guidance cue, each guidance cue $F^{(g)}$ is utilized to adapt the basic concentration parameter from $\boldsymbol{\alpha}_{0,k,g}$ to $\boldsymbol{\alpha}_{k,g}$:

$$\alpha_{k,g} = \alpha_{0,k,g} \cdot f_{\text{MLP}}\Big(\text{GAP}(\mathbf{F}^{(g)})\Big), \quad \boldsymbol{\alpha} \in \mathbb{R}^{B\times K\times G}. \tag{5}$$

where GAP is the global average pooling along the width and height, g is the index of guidance features, and k is the mode number of Mixed Dirichlet.

Similarly, the basic mode confidence $\mathbf{p}_0$ are adapted based on guidance cues:

$$\mathbf{q} = \text{softmax}\big(f_q(\text{Concat}_{\mathbf{g=1}}^{\mathbf{G}}\text{GAP}(\mathbf{F}^{(\mathbf{g})}))\big), \quad \mathbf{p} = \text{Norm}(\mathbf{p}_0 \odot \mathbf{q}). \tag{6}$$

where $f_q(\cdot)$ denotes an MLP that outputs guidance-aware gating weights over the K mixture modes.

Hence, GCDS could sample N candidate reliability strategies $\{\mathbf{W}^{(n)}\}_{n=1}^{N}$ to help dynamically generate multiple candidate predictions $Y_{\text{pred}}^{(n)}$ during the

training stage. Then, inspired by the core thinking of quantifying the relative advantages among different candidate tokens, we propose to encourage the segmentation model in learning optimal lesion prediction upon candidate reliability strategies $\{\mathbf{W}^{(n)}\}_{n=1}^{N}$.

2.3 Structure Advantage Learning

Considering that weak labels may lack sufficient supervision around lesion boundaries and further cause boundary ambiguity, we propose to encourage the segmentation model to autonomously explore an optimal lesion structure. We empirically observed that the VLM semantic scores have a consistent positive correlation with annotation accuracy.

Spatial Quality Score: To ensure basic lesion localization consistency, given the prediction Y_{pred}^{n} under the n-th GCDS candidate strategy $\mathbf{W}^{(n)}$, we leverage the sparse foreground and background point sets P^{+} and P^{-} based on the VLM-suggested mask Y_c to compute a Spatial Quality Score $R_S^n \in [0,1]$:

$$R_S^n = \frac{1}{|P^{+}|} \sum_{(u,v)\in P^{+}} P_{\text{pred}}^{n}(u,v) + \frac{1}{|P^{-}|} \sum_{(u,v)\in P^{-}} \big(1 - P_{\text{pred}}^{n}(u,v)\big). \tag{7}$$

This score evaluates whether a candidate prediction aligns with the limited but reliable point-level supervision, without enforcing dense pixel-level constraints.

Temporal Semantic Score: While spatial cues ensure coarse localization, accurate lesion segmentation in ultrasound videos requires consistent structural reasoning across time. To further encourage optimal lesion prediction and enhance the penalty for inaccurate lesion structures, we employ the VLM [22] as a semantic evaluator to perform relative comparison among candidate predictions. For each frame I_t, the VLM compares the set of candidate predictions $\{Y_{pred,t}^{n}\}_{n=1}^{N}$ and assigns semantic scores s_t^n with corresponding reasoning r_t^n.

We then aggregate s_t^n and r_t^n along the temporal dimension for each candidate prediction, and then prompt the VLM to further summarize and provide a global assessment as the Temporal Semantic Score $R_T^n \in [0,1]$:

$$\{(s_t^n, r_t^n)\}_{n=1}^{N} = \Psi_{\text{VLM}}(I_t, \{Y_{t,\text{pred}}^{j}\}_{j=1}^{N}), \quad R_T^n = \Phi_{\text{VLM}}\big(\text{Cat}_{t=1}^{T}\{(s_t^n, r_t^n)\}\big). \tag{8}$$

Hence, the VLM could be used for relative quality ranking across multiple candidate predictions and encourage the model to generate optimal masks.

Relative Quality Quantify: To quantify the relative quality of these candidate predictions under different strategies and select an optimal prediction as the final output, we combine the spatial and temporal evaluation scores (R_S^n and R_T^n) to obtain an overall quality score R^n, which is then normalized across the N strategies to compute the relative quality weight A^n:

$$R^n = R_S^n + R_T^n, \quad A^n = \frac{R^n - \mu(R)}{\sigma(R) + \epsilon}. \tag{9}$$

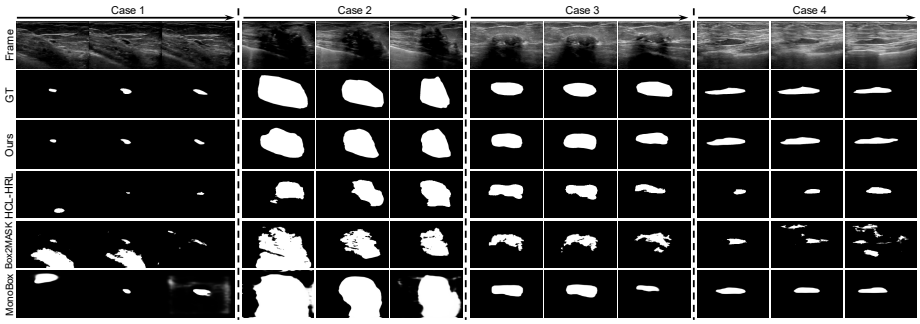

Fig. 2. Visual comparisons of our proposed method and recent SOTA methods, including small lesions, blurred boundaries, motion artifacts, and irregular shapes.

where $\mu(R)$ and $\sigma(R)$ denote the mean and standard deviation across candidate scores, respectively. This formulation encourages the model to favor structurally superior predictions while suppressing inferior ones.

Loss Optimization: Each candidate prediction is formulated as a pixel-wise model predicted probability map $\mathbf{P}^n_{\text{pred},t}$ for the n-th GCDS strategy at frame I_t. Here, we adopt a stop-gradient pseudo target $\hat{Y}^n_t = stopgrad(Y^n_t)$ and encourage a higher-quality candidate predictions exert stronger influence during training [16]:

$$\log \pi_\theta\Big(\hat{\mathbf{Y}}^n_t \mid I_t, n\Big) = \hat{Y}^n_t \log P^n_{\text{pred},t} + (1-\hat{Y}^n_t)\log(1-P^n_{\text{pred},t}). \tag{10}$$

Finally, the overall training loss $\mathcal{L}_{total}$ of our method can be formulated as:

$$\mathcal{L}_{total} = -\frac{1}{TNHW}\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{h=1}^{H}\sum_{w=1}^{W}\log \pi_\theta\Big(\hat{\mathbf{Y}}^n_t \mid I_t, n\Big) \times A^n + PBCE(Y_{pred}, Y_c). \tag{11}$$

where $PBCE$ is the point-level BCE loss between the coarse pseudo-mask Y_c and the most confident prediction Y_{pred}, T is the number of frames, and n indexes n-th sampled candidate strategy.

3 Experiments

3.1 Implementation

We follow the original dataset settings to conduct experiments on the DPSTT [9], ERUS-10K [7], and our collected ULBV400+, which contains 433 breast lesion ultrasound videos with a spatial resolution of 510×470. In our experiments, we deploy a public diffusion segmentation model [25] as the backbone on 2 RTX 4090 GPUs. Then Qwen3-VL-8B [22] is used as the VLM supervisor during training. The batch size is set to 4, the learning rate to 1e-4. All input frames are resized to 256×256. No additional data augmentation is applied except for the data

Table 1. Quantitative comparison with different methods on our ULBV400+ dataset.

Methods	Jaccard	Dice	Precision	Recall
FSRM [26]	56.65±2.65	68.03±2.58	82.09±2.35	68.34±2.88
BBTP [20]	50.05±2.63	61.93±2.77	81.82±2.37	60.32±3.05
WeakMedSAM [17]	50.64±2.76	62.02±2.88	80.74±3.03	57.79±3.12
HCL-HRL [3]	51.67±2.78	62.86±2.92	77.78±2.54	66.67±3.37
MonoBox [6]	59.06±2.74	70.48±2.47	78.23±2.69	73.82±2.59
Box2Mask [13]	56.36±2.54	68.13±2.47	77.48±2.80	70.00±2.33
Ours	**68.26±4.17**	**78.90±3.53**	**85.21±4.16**	**78.61±4.22**

Table 2. Quantitative comparison with different methods on ERUS-10K [7] dataset.

Methods	Jaccard	Dice	Precision	Recall
FSRM [26]	56.47±2.32	68.91±2.25	75.71±2.18	72.08±2.58
BBTP [20]	52.51±2.59	64.42±2.67	77.73±2.32	62.76±2.97
WeakMedSAM [17]	47.53±2.41	60.42±2.49	69.07±2.55	58.61±2.62
HCL-HRL [3]	48.84±2.10	62.61±2.17	69.03±2.21	63.82±2.69
MonoBox [6]	54.77±2.04	68.03±2.11	72.77±2.21	69.67±2.49
Box2Mask [13]	58.47±1.89	71.69±1.79	72.00±1.93	75.79±1.85
Ours	**63.08±2.88**	**75.87±2.09**	**79.67±2.14**	**75.83±3.36**

normalization. We compare our method with recent SOTA methods using their official open-source code [3, 6, 13, 17, 20, 26]. Manual correction is applied only to the BBoxes whose confidence scores are below 7, while the BBoxes provided by the VLM are directly used for model training for all remaining frames. For stable end-to-end training, the Dirichlet distribution in the GCDS is sampled using a reparameterized Gamma-based implementation, enabling pathwise gradient backpropagation through the sampled guidance weights. Each frame was independently annotated 3 times, and GT was generated via pixel-wise majority voting to reduce observer variation and enhance reliability.

3.2 Experimental Results

Table 1, Table 2, and Table 3 quantitatively report the experimental results on our ULBV400+ dataset, the ERUS-10K dataset [7], and the DPSTT dataset [9]. We can find that our method consistently achieves better performance than recent SOTAs. Our method is compatible with various VLMs. As shown in Table 5, during the training process, we encourage the Qwen3 [22] to generate a BBox with its confidence score ranging from 0 to 10 for each frame. Compared to recent methods, our method could save clinicians around 60% of the manual annotation workload on our ULBV400+ dataset, the ERUS-10K dataset [7], and the DPSTT dataset [9]. As shown in Fig. 2, we can find that our method could achieve better segmentation results and is more consistent with the GT annotation.

Table 3. Quantitative comparison with different methods on DPSTT [9] dataset.

Methods	Jaccard	Dice	Precision	Recall
FSRM [26]	63.23±5.48	77.36±4.19	84.24±5.98	71.99±7.10
BBTP [20]	60.95±5.33	75.63±4.20	85.30±1.87	68.13±6.41
WeakMedSAM [17]	58.67±6.83	70.06±6.47	81.44±5.83	66.57±6.42
HCL-HRL [3]	61.68±5.17	72.81±5.29	76.39±5.69	77.24±8.61
MonoBox [6]	65.01±4.62	75.77±4.58	84.30±7.12	73.29±2.98
Box2Mask [13]	65.25±3.99	75.81±3.59	80.48±4.74	76.34±3.36
Ours	**72.30±2.31**	**81.66±2.65**	**85.94±3.78**	**81.01±2.12**

Table 4. Ablation study for proposed modules on DPSTT.

Methods	Jaccard	Dice	Precision	Recall
M1	65.39±4.63	75.94±4.34	84.92±3.82	72.99±6.32
M2	67.19±4.25	77.12±4.52	84.97±3.79	74.67±6.52
M3	67.77±3.20	77.97±3.43	83.44±3.76	77.24±4.54
M4	70.80±2.88	80.24±3.11	85.33±3.73	79.31±2.94
Ours	**72.30±2.31**	**81.66±2.65**	**85.94±3.78**	**81.01±2.12**

Table 5. Percentage of manually labeled BBoxes on US dataset.

Mannal Percentage Supervision		DPSTT	ASTR	ULBV400+
BBTP [20]	BBox	100%	100%	100%
FSRM [26]	BBox	100%	100%	100%
MonoBox [6]	BBox	100%	100%	100%
Box2Mask [13]	BBox	100%	100%	100%
Ours	BBox	**32.59%**	**27.51%**	**39.98%**

3.3 Ablation Studies

As shown in Table 4, we conduct ablation experiments to evaluate the effectiveness of our proposed modules. We consider several baseline networks: 1) M1 is constructed by only using the normal spatial domain frame feature and training with normal point-level dice and BCE loss. 2) M2 is constructed by further incorporating high-frequency wavelet component frame features [1] on M1. 3) M3 further incorporates the BiomedCLIP-embeded [24] VLM knowledge as additional guidance. 4) M4 is further training using our proposed Structure Advantage Loss without the Temporal Semantic Score R_T. We can observe that wavelet-domain features and VLM-derived knowledge could consistently improve the segmentation performance. Our proposed Structure Advantage Learning further enhances the segmentation performance compared to baseline models.

4 Conclusion

In this paper, we propose a large-scale ultrasound VOS dataset with 433 videos and a weakly supervised method that leverages expert knowledge from a frozen VLM to substantially reduce annotation cost. We propose the GCDS diffusion strategy with the Structure Advantage Learning that enables the segmentation model to progressively explore an optimal prediction mask. Extensive experiments on 3 datasets show consistent improvements over recent state-of-the-art methods. We believe this work shows the potential of leveraging foundation models to reduce annotation burden for efficient medical video analysis.

Acknowledgment. This work was supported by Guangdong Science and Technology Department (2024ZDZX2004), the Natural Science Foundation of China under Grant U24A20278, and Shenzhen Science and Technology Program (JCYJ20241202152803005, NO.KJZD20240903102717023).

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

References

1. Bentley, P.M., McDonnell, J.: Wavelet transforms: an introduction. Electronics & communication engineering journal **6**(4), 175–186 (1994)
2. Chen, R., Su, T., Wang, H.: Wavecl: Wavelet calibration learning for referring video object segmentation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3856–3864 (2025)
3. Chi, J., Li, Z., Lin, G., Sun, M., Yu, X.: Weakly supervised segmentation framework for thyroid nodule based on high-confidence labels and high-rationality losses. arXiv preprint arXiv:2502.19707 (2025)
4. Darroch, J., Ratcliff, D.: A characterization of the dirichlet distribution. Journal of the American Statistical Association **66**(335), 641–643 (1971)
5. Hao, P., Wang, H., Li, S., Xing, Z., Yang, G., Wu, K., Zhu, L.: Surgical-mamballm: Mamba2-enhanced multimodal large language model for vqla in robotic surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 573–583. Springer (2025)
6. Hu, Q., Yi, Z., Zhou, Y., Huang, F., Liu, M., Li, Q., Wang, Z.: Monobox: Tightness-free box-supervised polyp segmentation using monotonicity constraint. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3572–3580 (2025)
7. Jiang, Y., Hu, Y., Zhang, Z., Wei, J., Feng, C.M., Tang, X., Wan, X., Liu, Y., Cui, S., Li, Z.: Towards a benchmark for colorectal cancer segmentation in endorectal ultrasound videos: Dataset and model development. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 732–742. Springer (2024)
8. Li, H., Nourkhiz Mahjoub, H., Chalaki, B., Tadiparthi, V., Lee, K., Moradi Pari, E., Lewis, C., Sycara, K.: Language grounded multi-agent reinforcement learning with human-interpretable communication. Advances in Neural Information Processing Systems **37**, 87908–87933 (2024)
9. Li, J., Zheng, Q., Li, M., Liu, P., Wang, Q., Sun, L., Zhu, L.: Rethinking breast lesion segmentation in ultrasound: A new video dataset and a baseline network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 391–400. Springer (2022)
10. Li, J., Zhu, L., Shen, G., Zhao, B., Hu, Y., Zhang, H., Wang, W., Wang, Q.: Liver lesion segmentation in ultrasound: A benchmark and a baseline network. Computerized Medical Imaging and Graphics **123**, 102523 (2025)
11. Li, J., Zhu, L., Xing, Z., Zhao, B., Hu, Y., Lv, F., Wang, Q.: Cascaded inner-outer clip retformer for ultrasound video object segmentation. IEEE Journal of Biomedical and Health Informatics (2024)
12. Li, S., Yang, Y., Xing, Z., Wang, H., Hao, P., Li, X., Liu, Z., Zhang, Q., Zhu, L.: Synerdetect: Hierarchical synergistic learning for generalizable ai-generated image detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 6405–6414 (2026). https://doi.org/10.1609/aaai.v40i8.37568
13. Li, W., Liu, W., Zhu, J., Cui, M., Yu, R., Hua, X., Zhang, L.: Box2mask: Box-supervised instance segmentation via level-set evolution. IEEE transactions on pattern analysis and machine intelligence **46**(7), 5157–5173 (2024)

14. Liao, G., Jogan, M., Koushik, S., Eaton, E., Hashimoto, D.A.: Disentangling spatio-temporal knowledge for weakly supervised object detection and segmentation in surgical video. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 8013–8023. IEEE (2025)
15. Ng, K.W., Tian, G.L., Tang, M.L.: Dirichlet and related distributions: Theory, methods and applications (2011)
16. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025)
17. Wang, H., Huai, L., Li, W., Qi, L., Jiang, X., Shi, Y.: Weakmedsam: Weakly-supervised medical image segmentation via sam with sub-class exploration and prompt affinity mining. IEEE Transactions on Medical Imaging (2025)
18. Wang, H., Chen, W., Luo, X., Xing, Z., Liu, L., Qin, J., Wu, S., Zhu, L.: Toward fair and accurate cross-domain medical image segmentation: A vlm-driven active domain adaptation paradigm. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24102–24112 (2025)
19. Wang, H., Chen, Y., Chen, W., Xu, H., Zhao, H., Sheng, B., Fu, H., Yang, G., Zhu, L.: Serp-mamba: Advancing high-resolution retinal vessel segmentation with selective state-space model. IEEE Transactions on Medical Imaging (2025)
20. Wang, J., Xia, B.: Bounding box tightness prior for weakly supervised image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 526–536. Springer (2021)
21. Xu, H., Nie, Y., Wang, H., Chen, Y., Li, W., Ning, J., Liu, L., Wang, H., Zhu, L., Liu, J., et al.: Medground-r1: Advancing medical image grounding via spatial-semantic rewarded group relative policy optimization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 391–401. Springer (2025)
22. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
23. Zhang, D., Han, J., Cheng, G., Yang, M.H.: Weakly supervised object localization and detection: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(9), 5866–5885 (2021)
24. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
25. Zhou, H., Wang, H., Ye, T., Xing, Z., Ma, J., Li, P., Wang, Q., Zhu, L.: Timeline and boundary guided diffusion network for video shadow detection. In: ACM Multimedia 2024 (2024)
26. Zhu, Z., Shi, J., Zhao, M., Wang, Z., Qiao, L., An, H.: Contrast learning based robust framework for weakly supervised medical image segmentation with coarse bounding box annotations. In: International Workshop on Computational Mathematics Modeling in Cancer Analysis. pp. 110–119. Springer (2023)