DualProtoSeg: Simple and Efficient Design with Text- and Image-Guided Prototype Learning for Weakly Supervised Histopathology Image Segmentation

Anh M. Vu[*1] MVU9@COUGARNET.UH.EDU
Khang P. Le[*2] KHANG.LEKHANGLE23@HMCUT.EDU.VN
Trang T. K. Vo[*3] TRANGVTK.18@GRAD.UIT.EDU.VN
Ha Thach [4] NGUYENBICHHA.THACH@STUDENT.UTS.EDU.AU
Huy Hung Nguyen[5] 25HUNG.NH@VINUNI.EDU.VN
David Yang[6] DONGJUN.YANG@EMORY.EDU
Han H. Huynh[7] M658112001@TMU.EDU.TW
Quynh Nguyen[1] QTNGUY50@COUGARNET.UH.EDU
Tuan M. Pham[1] TMPHAM42@COUGARNET.UH.EDU
Tuan-Anh Le[1] TLE99@COUGARNET.UH.EDU
Minh H. N. Le[8] JOHNMINHLE@IEEE.ORG
Thanh-Huy Nguyen[9] THANHHUN@ANDREW.CMU.EDU
Akash Awasthi[1] AAWASTH3@COUGARNET.UH.EDU
Chandra Mohan[1] CMOHAN@CENTRAL.UH.EDU
Zhu Han[1] ZHAN2@CENTRAL.UH.EDU
Hien Van Nguyen[1] HVNGUY35@CENTRAL.UH.EDU

[1] *University of Houston, Houston, TX, USA*
[2] *Ho Chi Minh City University of Technology, Vietnam*
[3] *University of Information Technology, Ho Chi Minh City, Vietnam*
[4] *University of Technology Sydney, Australia*
[5] *Vin University, Hanoi, Vietnam*
[6] *Department of Computer Science, Emory University, Atlanta, GA, USA*
[7] *College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan*
[8] *Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, NY, USA*
[9] *School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA*

Editors: Under Review for MIDL 2026

[*] Contributed equally

Abstract

Weakly supervised semantic segmentation (WSSS) in histopathology seeks to reduce annotation cost by learning from image-level labels, yet it remains limited by inter-class homogeneity, intra-class heterogeneity, and the region-shrinkage effect of CAM-based supervision. We propose a simple and effective prototype-driven framework that leverages vision–language alignment to improve region discovery under weak supervision. Our method integrates CoOp-style learnable prompt tuning to generate text-based prototypes and combines them with learnable image prototypes, forming a dual-modal prototype bank that captures both semantic and appearance cues. To address oversmoothing in ViT representations, we incorporate a multi-scale pyramid module that enhances spatial precision and improves localization quality. Experiments on the BCSS-WSSS benchmark show that our approach surpasses existing state-of-the-art methods, and detailed analyses demonstrate the benefits of text description diversity, context length, and the complementary behavior of text and image prototypes. These results highlight the effectiveness of jointly leveraging textual semantics and visual prototype learning for WSSS in digital pathology. Code is available at: https://github.com/maianhpuco/DualProtoSeg.git

Keywords: Weakly supervised semantic segmentation, Histopathology, Prototype Learning.

1. Introduction

Weakly supervised segmentation (WSS) has become a practical alternative to costly pixel-level annotation in histopathology (Liu et al., 2021; Minaee et al., 2022), relying instead on weak labels such as image-level tags (Zhou, 2017). Most WSS follows a two-stage paradigm: a classifier trained on weak labels generates pseudo-masks (e.g., CAMs or instance labels), which then supervise a fully supervised segmentation model. CAMEL (Xu et al., 2019), for instance, converts MIL-derived instance labels into approximate pixel masks for downstream segmentation. Image-level WSSS methods such as TransWS (Zhang et al., 2022) use a classification backbone to produce coarse localization cues and refine them through a transformer-based decoder, while TPRO (Zhang et al., 2023) leverages class descriptions as text prompts and a knowledge-attention module to build pseudo-masks. Other weak annotation strategies include sparse point methods (Qu et al., 2020), which expand point labels into coarse masks and refine boundaries via CRF (Krähenbühl and Koltun, 2011), and bounding-box–based methods (Kervadec et al., 2020), which enforce tightness and background constraints directly in the loss function, enabling training with only box labels.

Class-activation-based (Grad-CAM-style) WSSS remains one of the most widely used paradigms: a classification network is trained with image-level labels, and its gradient-based activation maps are used as pseudo-masks for segmentation. However, CAMs inherently highlight only the most discriminative regions that support the image-level prediction, rather than representing the full spatial extent of each tissue type. This limitation is particularly problematic in histopathology, where intra-class heterogeneity (large morphological variation within a class) and inter-class homogeneity (high visual similarity between different classes) are pervasive. As a result, CAMs tend to collapse onto a small subset of visually distinctive patterns and systematically miss large portions of relevant structures. To compensate for these weaknesses, prior works have proposed multi-scale CAM fusion (Ma et al., 2020), instance- or region-aware activation mechanisms (Torabi et al., 2025), and additional priors such as self-supervision, background suppression, saliency, and vision–language guidance (Feng et al., 2025). While these strategies improve coverage and bound-

ary quality, they still operate within the inherent constraints of activation-based localization, and thus remain limited in their ability to model the full morphological diversity required for reliable histopathology WSSS. Recent attribution methods such as CIG (Vu et al., 2025) show that stronger class-discriminative cues exist beyond CAMs, but they remain interpretability tools rather than sources of dense masks, reinforcing the need for richer representation mechanisms.

Prototype-based WSSS offers a promising remedy to the limitations of CAM-based methods because prototypes capture characteristic morphological patterns rather than relying solely on highly discriminative regions. By modeling multiple representative appearances per class, prototype systems naturally mitigate intra-class heterogeneity and reduce confusion arising from inter-class homogeneity. Clustering-based variants such as PBIP (Tang et al., 2025) and SIPE (Chen et al., 2022) estimate semantic centers using K-means or affinity propagation and improve region coverage through contrastive prototype matching. Learnable prototype approaches, including ProtoPNet (Chen et al., 2019), ViLa-MIL (Shi et al., 2025), PIP-Net (Nauta et al., 2023), and LDP (Le et al., 2025), extend this idea by learning prototypes directly from data. Among them, LDP is a fully learnable prototype–based WSSS framework related to ours but fundamentally orthogonal, as it focuses on prototype refinement within a visual-only setting, whereas our work leverages text-image alignment to introduce multimodal semantic cues. Recent developments further enhance prototype robustness: (Xu et al., 2024b) aligns visual prototypes with text-derived semantics, ProtoSeg (Sacha et al., 2025) introduces a diversity constraint to prevent prototype collapse, and HisynSeg (Fang et al., 2025) enforces representational diversity via synthetic mixing to improve pseudo-mask completeness.

Despite these advances, prototype-based WSSS still faces key limitations. (1) Clustering-based methods introduce substantial computational overhead and are sensitive to hyperparameters. (2) Learnable prototype frameworks rarely leverage pathology-specific vision–language pretraining, missing an opportunity to incorporate rich semantic cues from paired image–text datasets. (3) Most prior work relies solely on visual prototypes, lacking the complementary semantic grounding that text–image alignment can provide. These gaps highlight the need for a more efficient, multimodal prototype framework that is robust to morphological variability and capable of leveraging modern pathology foundation models.

Modern histopathology encoders generally fall into two categories: vision–language models trained on large-scale pathology image–text pairs, and vision-only ViTs trained with powerful self-supervised objectives. Multimodal models such as CONCH (Lu et al., 2023) and QuiltNet capture pathology-specific image–text alignment and demonstrate strong zero-shot generalization, while vision-only models like UNI (Chen et al., 2023), CTransPath (Wang et al., 2021), and GigaPath (Xu et al., 2024a) provide robust visual representations learned from extensive SSL pretraining. However, most downstream WSSS pipelines treat these encoders purely as visual feature extractors and overlook both their multimodal capabilities and their hierarchical intermediate representations.

Inspired by advances in prototype learning and the strong capabilities of pathology-focused vision–language models for weak supervision, we propose **DualProtoSeg**, a simple and efficient framework that unifies text-guided and image-guided prototype learning for clustering-free WSSS in histopathology. Whole-slide patches are processed by a CLIP-style backbone (e.g., CONCH) to extract multi-scale visual features, while class descriptions are encoded by the text encoder and

refined through learnable prompt tuning. The resulting image and text prototypes are projected across feature scales, and cosine similarity with spatial features produces hierarchical CAMs that are fused into dense pseudo-masks. A semantic alignment loss pulls visual prototypes toward the corresponding text embeddings, and a diversity regularizer prevents prototype collapse, improving spatial coverage and localization. This complementary multimodal design mitigates oversmoothing in ViT-based backbones and produces accurate segmentation masks under weak supervision.

Our main contributions are as follows:

1. We introduce a weakly supervised segmentation framework that leverages pretrained CLIP-based vision-language representations, using text–image alignment as an additional source of semantic guidance.
2. We propose a dual-prototype architecture that jointly models visual and textual concepts, enabling complementary appearance- and semantics-driven pseudo-mask generation.
3. We incorporate a learnable prompt module in the text branch, allowing class descriptions to adapt to the dataset and providing stronger multimodal supervision.

2. Method

Overview. Dual-ProtoSeg performs weakly supervised semantic segmentation by combining multi-scale visual features with dual-modal prototypes. The frozen CONCH ViT-B/16 (Lu et al., 2023) encoder provides intermediate image features that are refined into a multi-scale pyramid, while the text branch encodes class descriptions using learnable context tokens. Text and image prototypes are projected to each scale and compared with spatial features via cosine similarity to generate class activation maps (CAMs). The multi-scale CAMs are upsampled, fused into pseudo-masks, and supervised with image-level classification losses. A DenseCRF (Krähenbühl and Koltun, 2011) further refines the fused CAM at inference to produce the segmentation mask.

2.1. Multi-Scale Visual Feature Extraction and Refinement

Image Feature Extraction. We use the frozen CONCH ViT-B/16 encoder, pretrained on large-scale pathology image–text pairs, to extract multi-scale visual representations. The transformer naturally produces a hierarchy of hidden states with increasing semantic richness across depth. To capture this information, we extract intermediate hidden states from several representative layers (e.g., 2, 5, 8, 11) using forward hooks. Each feature map is reshaped into a 2D spatial grid and denoted as

$$\mathbf{F}_k \in \mathbb{R}^{B\times D_{\text{vit}}\times H'\times W'}, \quad k = 1, \ldots, 4,$$

where D_{vit} is the token embedding dimension and $H'\times W'$ corresponds to the patch-token grid.

Feature Refinement. Although semantically meaningful, ViT intermediate features may exhibit blurred boundaries and weak localization. We therefore apply a lightweight refinement module to each $\mathbf{F}_k$. The features are first projected to $D_{\text{ref}}/2$ channels via a 1×1 convolution with GroupNorm and SiLU, followed by a series of residual blocks with 3×3 convolutions that improve spatial coherence and suppress noise. A final 1×1 convolution expands the representation to D_{ref}

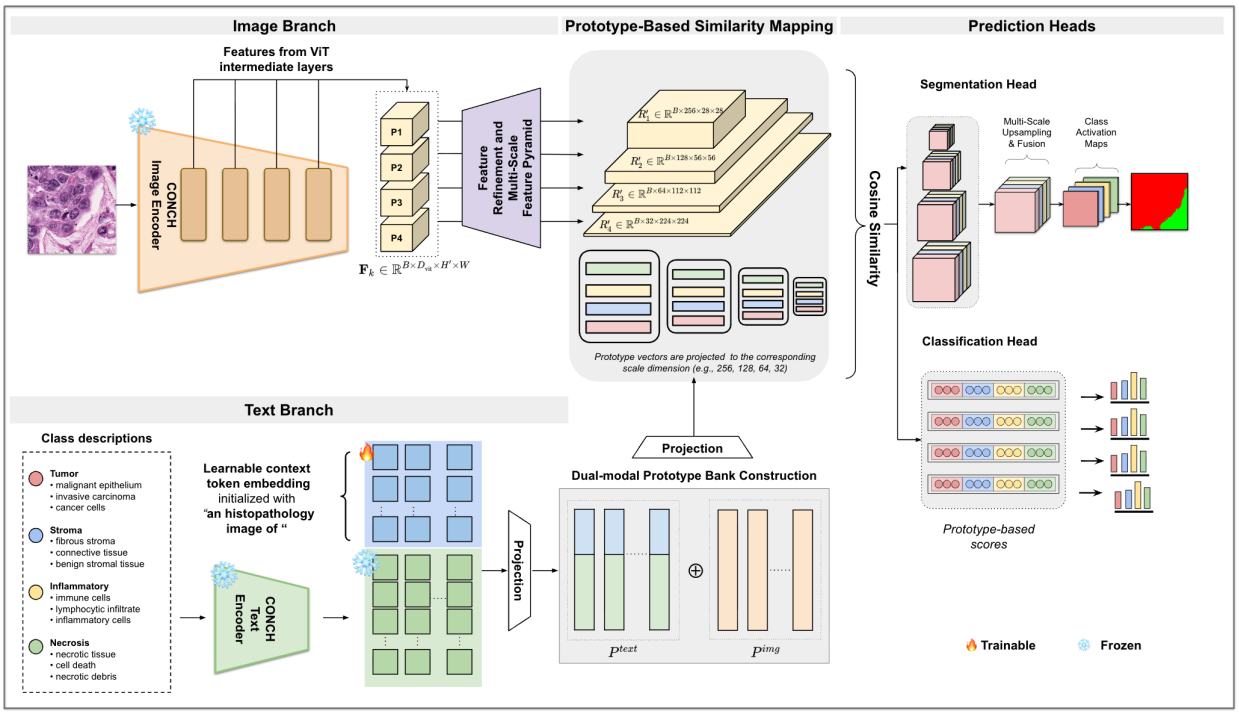

Figure 1: Overview of the Dual-ProtoSeg framework. Multi-scale image features (Image Branch) and prompt-guided text embeddings (Text Branch) form a dual-modal prototype bank, whose similarity to visual features generates multi-scale CAMs for weakly supervised segmentation (Prediction Heads).

channels and prepares it for prototype matching:

$$\mathbf{R}_k \in \mathbb{R}^{B\times D_{\text{ref}}\times H'\times W'}, \quad k = 1, \ldots, 4.$$

Multi-Scale Feature Pyramid. (Multi-ScaleFP) To integrate information at different spatial resolutions, the refined maps $\{\mathbf{R}_k\}$ are transformed into a pyramid of features with progressively increasing resolution and decreasing channel dimension. Each $\mathbf{R}_k$ is projected through a lightweight convolutional block and then resized by bilinear interpolation such that spatial dimensions double and channels halve at each level. Formally,

$$[R'_i]_{i=1}^{4} = \text{Multi-ScaleFP}([\mathbf{R}_i]_{i=1}^{4}), \quad R'_i \in \mathbb{R}^{B\times D_i\times H_i\times W_i},\ H_i = 2^i H',\ W_i = 2^i W',\ D_i = D_{\text{ref}}/2^{i-1}.$$

For example, if $\mathbf{R}_i$ has shape $[B, 512, 14, 14]$, the pyramid yields: $R'_1 : [B, 256, 28, 28]$, $R'_2 : [B, 128, 56, 56]$, $R'_3 : [B, 64, 112, 112]$, $R'_4 : [B, 32, 224, 224]$. This multi-scale construction produces spatially detailed and semantically consistent feature maps suitable for downstream dense prediction tasks.

2.2. Dual-Modal Prototype Bank Construction

Text-Based Prototypes with Learnable Context Tokens. To encode class-specific textual information, we adopt *Learnable Prompt Tokens* called ConchPromptLearner, inspired by CoOp

(Zhou et al., 2022), which allows adaptation of class descriptions while keeping the CONCH text encoder frozen. For C classes, each class c is assigned n_{ctx} learnable context tokens:

$$\mathbf{T}_c \in \mathbb{R}^{n_{\text{ctx}} \times D_{\text{text}}},$$

where D_{text} is the token embedding dimension of the CONCH text encoder. Each token sequence $\mathbf{T}_c$ is initialized from a base textual template (e.g., "a histopathology image of") and optionally perturbed with Gaussian noise: $\mathbf{T}_c^{(0)} = \mathbf{E}(\text{ctx\_init}) + \boldsymbol{\epsilon}_c, \quad \boldsymbol{\epsilon}_c \sim \mathcal{N}(0, \sigma^2)$, where $\mathbf{E}$ is the frozen token embedding. The tokens for all classes are stacked into a trainable parameter: $\mathbf{T}_{\text{ctx}} = \text{stack}(\mathbf{T}_1, \ldots, \mathbf{T}_C) \in \mathbb{R}^{C \times n_{\text{ctx}} \times D_{\text{text}}}$.

Each class c has n_{desc} textual descriptions $\{t_c^{(d)}\}$. For each description, the learnable context tokens $\mathbf{T}_c$ are inserted into the prompt (e.g., $[\text{BOS}, \mathbf{T}_c, t_c^{(d)}, \text{EOS}]$), which is padded or truncated to 77 tokens. The frozen text encoder processes the prompt, and the end-of-text (EOT) embedding is extracted, linearly projected, and L2-normalized to obtain a text prototype:

$$\mathbf{f}_{c,d}^{\text{text}} = \text{normalize}\big(\text{LN}(\mathbf{H}_{\text{EOT}})W_{\text{proj}}\big).$$

At this stage, each textual description produces a separate prototype vector, without averaging across descriptions. The collection of all text-based prototypes is

$$\mathbf{P}^{\text{text}} = \{\mathbf{f}_{c,d}^{\text{text}}\}_{c=1,\ldots,C;d=1,\ldots,n_{\text{desc}}} \in \mathbb{R}^{C \cdot n_{\text{desc}} \times D_{\text{text}}}.$$

During training, only the learnable tokens $\mathbf{T}_{\text{ctx}}$ and a lightweight visual projection layer are updated, while the CONCH backbone remains frozen. The ConchPromptLearner thus constructs a structured Text-based Prototype Bank, with each description forming an individual prototype aligned with visual features.

Learnable Image-based Prototypes. Let $\mathbf{P}^{\text{img}} \in \mathbb{R}^{C \times D_{\text{img}}}$ denote the set of learnable image prototypes, where D_{img} is the visual feature dimension (e.g., 512). These prototypes are initialized randomly and optimized during training to complement the text prototypes. Learnable image prototypes are also normalized:

$$\mathbf{P}_{\text{norm}}^{\text{img}} = \text{normalize}(\mathbf{P}^{\text{img}}) \in \mathbb{R}^{C \times D_{\text{img}}}.$$

Dual-modal Prototype Bank Construction. Text and image prototypes are then interleaved per class to form a combined prototype bank:

$$\mathbf{P}^{\text{combined}} = [\mathbf{P}_{\text{proj},1}^{\text{text}}, \mathbf{P}_{\text{norm},1}^{\text{img}}, \ldots, \mathbf{P}_{\text{proj},C}^{\text{text}}, \mathbf{P}_{\text{norm},C}^{\text{img}}] \in \mathbb{R}^{2C \times D_{\text{img}}}.$$

Finally, to match the multi-scale visual feature dimensions, the combined prototypes are projected through a series of learnable linear layers: $\mathbf{P}^{(s)} = \text{normalize}\big(\mathbf{P}^{\text{combined}} W_{\text{proto}}^{(s)}\big), \quad s = 1, \ldots, S$, where S is the number of feature scales (e.g., $S = 4$) and $W_{\text{proto}}^{(s)}$ projects to the corresponding scale dimension (e.g., 256, 128, 64, 32). This produces a multi-scale dual-modal prototype bank that jointly represents textual and visual information for each class and is used for attention-based or similarity-based reasoning with image features.

2.3. Multi-Scale Pseudo-Mask and Segmentation Mask Generation

Using the multi-scale dual-modal prototype bank and pyramid features $\{R'_i\}_{i=1}^4$, we generate class activation maps (CAMs) by computing the similarity between each spatial feature and the corresponding projected prototypes at each pyramid level.

For the i-th pyramid level, let $R'_i \in \mathbb{R}^{B\times D_i\times H_i\times W_i}$ be the feature map and $\mathbf{P}^{(i)} \in \mathbb{R}^{K\times D_i}$ the associated prototype bank. We compute cosine similarity with a learnable scaling factor logit_scale_i: $\text{CAM}_i = \text{logit\_scale}_i \cdot \text{normalize}(R'_i) \cdot \text{normalize}(\mathbf{P}^{(i)})^\top \in \mathbb{R}^{B\times K\times H_i\times W_i}$.

This is applied independently at all pyramid levels, producing multi-scale CAMs $\{\text{CAM}_i\}_{i=1}^4$. Each CAM is then upsampled to the original image size $(H_{\text{img}}, W_{\text{img}})$:

$$\widetilde{\text{CAM}}_i = \text{Upsample}(\text{CAM}_i, \text{size} = (H_{\text{img}}, W_{\text{img}})), \quad i = 1, \ldots, 4.$$

The upsampled CAMs are fused via element-wise averaging to produce the final pseudo-mask:

$$\text{CAM}_{\text{fused}} = \frac{1}{4}\sum_{i=1}^{4} \widetilde{\text{CAM}}_i \in \mathbb{R}^{B\times K\times H_{\text{img}}\times W_{\text{img}}}.$$

During training, we compute multi-label classification losses at each pyramid level by comparing pooled CAMs with image-level labels. The overall loss is a weighted sum of the four levels: $\text{cls\_loss} = \lambda_1\,\text{loss}_1 + \lambda_2\,\text{loss}_2 + \lambda_3\,\text{loss}_3 + \lambda_4\,\text{loss}_4$, which encourages the model to learn discriminative features at multiple resolutions for weakly supervised segmentation.

To generate the final segmentation mask, we apply post-processing methods that refine boundaries. In particular, we adopt the fully-connected DenseCRF approach from (Krähenbühl and Koltun, 2011) at inference, which enhances spatial coherence and sharpness of the pseudo-mask, producing a high-quality segmentation output.

3. Experments and Results

Experiment Settings. We evaluate on the BCSS-WSSS dataset with four classes (TUM, STR, LYM, NEC) using a CONCH backbone. Training uses AdamW (lr 1×10^{-5}, weight decay 0.001), batch size 64, and 20 epochs. The best validation checkpoint is used for testing, and performance is measured with mIoU and Dice scores.

Quantitative Results. On the BCSS-WSSS dataset (Table 1), our method achieves 71.35% mIoU and 83.14% mDice, outperforming the previous state-of-the-art (PBIP) by +1.93% and +1.30%, respectively (Table 1). It sets new best results for tumor (81.34% IoU), stroma (68.84% IoU), and necrosis (69.83% IoU), with improvements of up to +4.16% on these clinically critical classes, while achieving the highest lymphocyte Dice (79.08%). These consistent gains under extremely sparse supervision establish a new state-of-the-art on BCSS-WSSS and enable substantially more accurate delineation of key pathological structures in breast cancer histopathology.

Qualitative Results. Figure 2 presents qualitative segmentation results on representative BCSS-WSSS patches. While all methods capture the general tissue layout, our approach produces noticeably sharper boundaries and clearer structural details that better match the ground truth. Overall, our method delivers more accurate and consistent segmentation of tumor, stroma, necrosis,

Table 1: Segmentation results on **BCSS-WSSS**

Method	Metrics (%)		Per-class IoU (%)				Per-class Dice (%)			
	mIoU	mDice	TUM	STR	LYM	NEC	TUM	STR	LYM	NEC
TPRO	65.54	78.93	77.29	66.83	56.81	61.23	87.19	80.12	72.46	75.95
MLPS	61.58	75.95	72.98	62.58	52.03	58.73	84.38	76.99	68.45	74.00
Proto2Seg	57.42	72.24	63.25	58.28	53.27	54.89	77.49	73.64	67.78	70.08
PBIP	69.42	81.84	77.92	64.68	65.40	69.69	87.59	78.56	79.08	82.14
OURS	**71.35**	**83.14**	**81.34**	**68.84**	65.39	**69.83**	**89.71**	**81.54**	**79.08**	**82.23**

and lymphocyte regions, reflecting the quantitative improvements reported in Table 1. These results highlight the ability of our approach to generate precise and reliable segmentations under limited weak supervision.

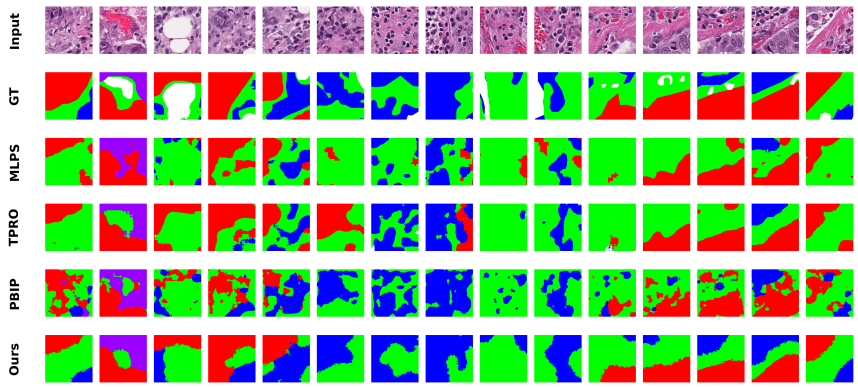

Figure 2: Qualitative results on BCSS-WSSS test patches. GT denotes ground truth segmentation

4. Ablation Studies and Additional Analysis

4.1. Complementarity of Text- and Image-based Prototypes.

Figure 3 shows that text prototypes activate on broad semantic regions, while image prototypes capture finer visual details recovering missed stroma areas (a, blue) and detecting subtle necrotic fragments in tumor (b, purple). Table 2 confirms this: adding image prototypes improves performance from 71.13% to 71.35% mIoU and yields the largest gains for lymphocytes. Image prototypes therefore provide complementary fine-grained cues that enhance segmentation beyond what text prototypes alone can offer.

4.2. Influence of Textual Descriptions and Context Length.

Table 3 shows that adding more textual descriptions per class steadily improves performance, as greater linguistic diversity strengthens text–image alignment. Using 10 descriptions achieves

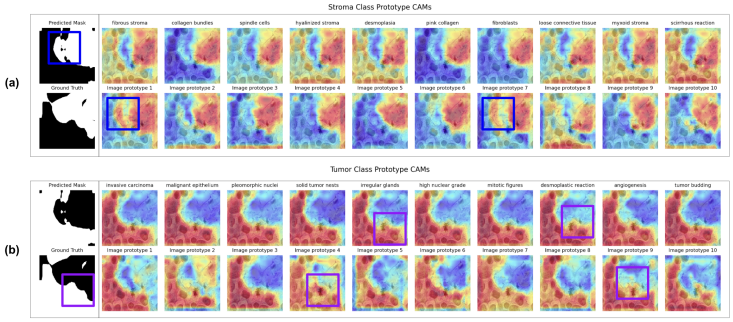

Figure 3: Complementary activation patterns of text-based (first row) and image-based (second row) prototypes. The first column shows the predicted mask (top) and the ground-truth mask (bottom). (a) Stroma: image prototypes recover missed regions (blue boxes). (b) Tumor: image prototypes detect fine-grained structures (purple boxes).

Table 2: Ablation of image prototypes in the hybrid prototype bank on **BCSS-WSSS** (10 text descriptions per class). "Yes" uses 10 additional image prototypes per class; "No" uses only the 10 text prototypes.

Image Proto.	Metrics (%)		Per-class IoU (%)				Per-class Dice (%)			
	mIoU	mDice	TUM	STR	LYM	NEC	TUM	STR	LYM	NEC
Yes	**71.35**	**83.14**	**81.34**	**68.84**	**65.39**	69.83	**89.71**	**81.54**	**79.08**	82.23
No	71.13	82.98	81.14	68.50	64.37	70.52	89.59	81.30	78.32	82.71

the best results (71.35% mIoU, 83.14% mDice), with notable gains for stroma and lymphocytes. Table 4 further shows that context length has minimal effect when only one description is used but becomes impactful with 10 descriptions, where longer contexts consistently improve accuracy and ctx=16 performs best. Overall, description diversity and prompt capacity contribute complementary benefits to segmentation quality.

Table 3: Ablation study on the number of textual descriptions per class (#Desc) on **BCSS-WSSS**

#Desc	Metrics (%)		Per-class IoU (%)				Per-class Dice (%)			
	mIoU	mDice	TUM	STR	LYM	NEC	TUM	STR	LYM	NEC
1	70.90	82.82	**81.36**	68.48	64.77	69.01	**89.72**	81.29	78.62	81.66
3	70.86	82.82	80.27	68.02	64.84	**70.32**	89.05	80.97	78.67	**82.57**
10	**71.35**	**83.14**	81.34	**68.84**	**65.39**	69.83	89.71	**81.54**	**79.08**	82.23

4.3. Pre-Training Assessment of Textual Descriptions.

Figure 4 analyzes textual description quality using zero-shot text–image retrieval AUC from the frozen CLIP text encoder, measuring how well each class's descriptions align with image features *before* training. Comparing these AUC values with final IoU (using 1, 3, and 10 descriptions)

Table 4: Ablation of textual description count and context length (ctx) on **BCSS-WSSS**.

ctx	Metrics (%)		Per-class IoU (%)				Per-class Dice (%)			
	mIoU	mDice	TUM	STR	LYM	NEC	TUM	STR	LYM	NEC
1 textual description per class										
4	70.81	82.76	**81.28**	**68.51**	**64.51**	68.96	89.67	**81.31**	78.43	81.63
8	70.35	82.44	80.46	68.43	63.65	68.85	89.17	81.26	77.79	81.55
16	**70.90**	**82.82**	81.36	68.48	64.77	**69.01**	**89.72**	81.29	**78.62**	**81.66**
10 textual descriptions per class										
4	70.55	82.58	80.62	68.26	63.40	**69.92**	89.27	81.14	77.60	**82.30**
8	70.73	82.69	81.27	68.47	63.31	69.87	89.67	81.28	77.54	82.26
16	**71.35**	**83.14**	**81.34**	**68.84**	**65.39**	69.83	**89.71**	**81.54**	**79.08**	82.23

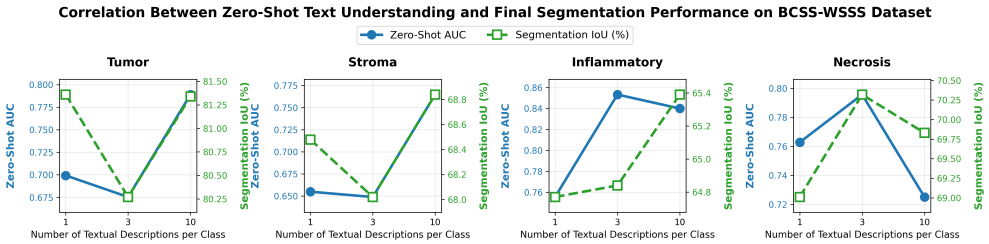

Figure 4: Relationship between zero-shot text–image alignment (AUC) and downstream segmentation IoU

shows that classes with strong initial alignment (tumor, inflammatory) perform well even with few descriptions, while weaker classes (stroma, necrosis) benefit markedly from more diverse descriptions. This suggests that zero-shot AUC is a practical pre-training diagnostic for identifying classes that require richer or more specific prompts.

5. Conclusion

We presented a weakly supervised segmentation framework that integrates CLIP-based vision–language features, a dual-prototype design, and a learnable prompt module to improve pseudo-mask quality in histopathology images. The combination of textual and visual prototypes provides complementary cues, enabling more complete coverage of tissue structures, while adaptive textual prompts strengthen alignment between descriptions and image features. Experiments on BCSS-WSSS show consistent improvements across classes and confirm the benefits of richer textual descriptions, longer prompt context, and the inclusion of image-based prototypes. These results demonstrate the effectiveness of multimodal supervision for enhancing weakly supervised segmentation.

Acknowledgments

This work was supported in part by the National Institutes of Health (NIH) under Grant 5R01DK134055-02.

References

Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition, 2019. URL https://arxiv.org/abs/1806.10574.

Qi Chen, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4278–4288. IEEE, June 2022. doi: 10.1109/cvpr52688.2022.00425. URL http://dx.doi.org/10.1109/CVPR52688.2022.00425.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H. Song, Muhammad Shaban, Mane Williams, Anurag Vaidya, Sharifa Sahai, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Walt Williams, Long Phi Le, Georg Gerber, and Faisal Mahmood. A general-purpose self-supervised model for computational pathology, 2023. URL https://arxiv.org/abs/2308.15474.

Zijie Fang, Yifeng Wang, Peizhang Xie, Zhi Wang, and Yongbing Zhang. Hisynseg: Weakly-supervised histopathological image segmentation via image-mixing synthesis and consistency regularization. *IEEE Transactions on Medical Imaging*, 44(4):1765–1782, April 2025. ISSN 1558-254X. doi: 10.1109/tmi.2024.3520129. URL http://dx.doi.org/10.1109/TMI.2024.3520129.

Siyang Feng, Hualong Zhang, Xianjing Zhao, Liting Shi, Zhenbing Liu, Rushi Lan, Lei Shi, and Xipeng Pan. Wave-aware weakly supervised histopathological tissue segmentation with cross-scale logits distillation. *IEEE Trans. Med. Imaging*, 2025. doi: 10.1109/TMI.2025.3637119. Advance online publication.

Hoel Kervadec, Jose Dolz, Shanshan Wang, Eric Granger, and Ismail Ben Ayed. Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision, 2020. URL https://arxiv.org/abs/2004.06816.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011.

Khang Le, Anh Mai Vu, Thi Kim Trang Vo, Ha Thach, Ngoc Bui Lam Quang, Thanh-Huy Nguyen, Minh H. N. Le, Zhu Han, Chandra Mohan, and Hien Van Nguyen. Lpd: Learnable prototypes with diversity regularization for weakly supervised histopathology segmentation, 2025. URL https://arxiv.org/abs/2512.05922.

Xiangbin Liu, Liping Song, Shuai Liu, and Yudong Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3), 2021. ISSN 2071-1050. doi: 10.3390/su13031224. URL https://www.mdpi.com/2071-1050/13/3/1224.

Ming Y. Lu, Bowen Chen, Andrew Zhang, Drew F. K. Williamson, Richard J. Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19764–19775, June 2023.

Xiao Ma, Zexuan Ji, Sijie Niu, Theodore Leng, Daniel L. Rubin, and Qiang Chen. Ms-cam: Multi-scale class activation maps for weakly-supervised segmentation of geographic atrophy lesions in sd-oct images. *IEEE Journal of Biomedical and Health Informatics*, 24(12):3443–3455, 2020. doi: 10.1109/JBHI.2020.2999588.

Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022. doi: 10.1109/TPAMI.2021.3059968.

Meike Nauta, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. PIP-Net: Patch-Based Intuitive Prototypes for Interpretable Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15509–15518, 2023.

Hui Qu, Pengxiang Wu, Qiaoying Huang, Jingru Yi, Zhennan Yan, Kang Li, Gregory M. Riedlinger, Subhajyoti De, Shaoting Zhang, and Dimitris N. Metaxas. Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. *IEEE Transactions on Medical Imaging*, 39(11):3655–3666, November 2020. ISSN 1558-254X. doi: 10.1109/tmi.2020.3002244. URL http://dx.doi.org/10.1109/TMI.2020.3002244.

Mikołaj Sacha, Dawid Rymarczyk, Łukasz Struski, Jacek Tabor, and Bartosz Zieliński. Protoseg: Interpretable semantic segmentation with prototypical parts, 2025. URL https://arxiv.org/abs/2301.12276.

Jiangbo Shi, Chen Li, Tieliang Gong, Yefeng Zheng, and Huazhu Fu. Vila-mil: Dual-scale vision-language multiple instance learning for whole slide image classification, 2025. URL https://arxiv.org/abs/2502.08391.

Qingchen Tang, Lei Fan, Maurice Pagnucco, and Yang Song. Prototype-based image prompting for weakly supervised histopathological image segmentation, 2025. URL https://arxiv.org/abs/2503.12068.

Ali Torabi, Sanjog Gaihre, MD Mahbubur Rahman, and Yaqoob Majeed. Localized region guidance for class activation mapping in wsss, 2025. URL https://arxiv.org/abs/2509.12496.

Anh Mai Vu, Tuan L. Vo, Ngoc Lam Quang Bui, Nam Nguyen Le Binh, Akash Awasthi, Huy Quoc Vo, Thanh-Huy Nguyen, Zhu Han, Chandra Mohan, and Hien Van Nguyen. Contrastive integrated gradients: A feature attribution-based method for explaining whole slide image classification, 2025. URL https://arxiv.org/abs/2511.08464.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image

classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 186–195. Springer, 2021.

Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. Camel: A weakly supervised learning framework for histopathology image segmentation, 2019. URL https://arxiv.org/abs/1908.10555.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 2024a.

Zhongxing Xu, Feilong Tang, Zhe Chen, Yingxue Su, Zhiyi Zhao, Ge Zhang, Jionglong Su, and Zongyuan Ge. Toward modality gap: Vision prototype learning for weakly-supervised semantic segmentation with clip, 2024b. URL https://arxiv.org/abs/2412.19650.

Shaoteng Zhang, Jianpeng Zhang, and Yong Xia. Transws: Transformer-based weakly supervised histology image segmentation. In *Machine Learning in Medical Imaging: 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, page 367–376, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-21013-6. doi: 10.1007/978-3-031-21014-3_38. URL https://doi.org/10.1007/978-3-031-21014-3_38.

Shaoteng Zhang, Jianpeng Zhang, Yutong Xie, and Yong Xia. Tpro: Text-prompting-based weakly supervised histopathology tissue segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023: 26th International Conference, Vancouver, BC, Canada, October 8–12, 2023, Proceedings, Part I*, page 109–118, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-43906-3. doi: 10.1007/978-3-031-43907-0_11. URL https://doi.org/10.1007/978-3-031-43907-0_11.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1): 44–53, 08 2017. ISSN 2095-5138. doi: 10.1093/nsr/nwx106. URL https://doi.org/10.1093/nsr/nwx106.