ProBAG: Prototype-Guided Boundary-Aware Graph Diffusion for Weakly Supervised Histopathology Segmentation

Duy-Dong Nguyen[1], Le-Van Thai[1], Hoai Nhan Pham[1], Ngoc Lam Quang Bui[2], Tam Tran[6(✉)[0009−0000−9702−425X]], and Zhi Huang[4(✉)]

[1] AI VIETNAM Lab, Vietnam
[2] Department of Mechanical System Engineering, Jeonbuk National University, Republic of Korea
[3] John T. Milliken Department of Medicine, Washington University School of Medicine, Saint Louis, MO, USA
tamt@wustl.edu
[4] Perelman School of Medicine, University of Pennsylvania, USA
zhi.huang@pennmedicine.upenn.edu

Abstract. Weakly supervised semantic segmentation enables histopathology tissue segmentation from image-level annotations, avoiding costly pixel-level labeling by expert pathologists. However, CAM-based methods often localize only highly discriminative regions and remain unreliable near tissue interfaces. We propose ProBAG, a stage-1 pseudo-mask generator that combines dataset-specific visual prototypes with pathology-aligned CONCH text prototypes over multi-scale frozen UNI features. ProBAG introduces two complementary mechanisms: class-wise power recalibration that reshapes inter-class competition while preserving the total foreground activation mass at each pixel, and one-step graph diffusion in which feature affinities are penalized by a late-transformer attention-context discrepancy used as a soft structural boundary cue. The resulting stage-1 pseudo-masks require neither CRF nor an external segmentation model; for complete two-stage comparison, they additionally supervise a downstream Phikon-FPN segmenter. Experiments on BCSS-WSSS and LUAD-HistoSeg show consistent gains over recent WSSS approaches, while ablations indicate that pathology-aligned text semantics provide the largest improvement and graph refinement provides a smaller complementary gain. The code is available at: https://github.com/wterrr/WSSS

Keywords: Weakly supervised semantic segmentation · Histopathology · Text-visual prototypes · Boundary-aware graph diffusion

1 Introduction

Histopathology image segmentation is a core task in computational pathology, supporting quantitative analysis of tumor, stroma, lymphocyte, and necrosis

regions in hematoxylin and eosin (H&E) images. Although deep learning has advanced digital pathology, dense semantic segmentation still requires costly pixel-level annotation by expert pathologists [3,4]. Weakly supervised semantic segmentation (WSSS) reduces this burden by learning pseudo-masks from cheaper annotations such as image-level tissue labels[16].

Most image-level WSSS methods rely on class activation maps (CAMs), which localize discriminative regions for classifier predictions [20,28]. Prior work has improved CAM quality through seed expansion, affinity propagation, inter-pixel relations, equivariance constraints, transformer attention, and class re-activation [1,2,6,12,19,24]. However, CAM supervision remains indirect, as classifiers often attend only to the most discriminative regions rather than capturing full object extent [13,26]. This limitation is particularly severe in histopathology, where tissues exhibit strong visual similarity and substantial intra-class morphological variation [21]. As a result, CAM-derived pseudo-masks are often incomplete, class-biased, and unreliable near boundaries.

Recent histopathology WSSS studies incorporate stronger domain priors to address these challenges. WSSS4LUAD establishes lung adenocarcinoma tissue segmentation from patch-level labels [10], while PistoSeg exploits dataset synthesis and feature consistency [8]. TPRO introduces text prompting for tissue segmentation [27], and prototype-based prompting leverages visual prototype banks to mitigate inter-class homogeneity and intra-class heterogeneity [21]. These methods show that semantic and prototype priors significantly improve CAM localization. However, two challenges remain: class-dependent activation imbalance in fused CAMs, where highly activated regions dominate foreground prediction, and the lack of boundary-aware refinement beyond feature-similarity propagation in prototype-based methods.

Existing prototype-based WSSS methods primarily improve semantic localization, but two failure modes remain insufficiently separated. First, independently normalized class CAMs can exhibit markedly different activation distributions, allowing broad or sharply activated tissues to dominate pixel-wise competition. Second, feature-only affinity propagation can leak across interfaces between adjacent tissues with similar local appearance. ProBAG targets these failures separately through foreground-mass-preserving class recalibration and attention-context-regularized graph propagation.

Recent advances in pathology foundation models and vision-language models provide useful cues for addressing these challenges. Frozen vision transformers can provide dense multi-scale representations, while pathology-aligned vision-language models offer semantically meaningful tissue embeddings [5,7,15]. These developments motivate a WSSS framework that jointly exploits visual prototypes, pathology text priors, and foundation-model attention cues. Recent studies have also demonstrated the utility of pathology foundation models for histopathology segmentation [9,22].

Accordingly, our contribution lies not in introducing prototypes or graph propagation in isolation, but in coordinating pathology-aligned semantic pro-

totypes with two explicit corrections for class competition and cross-interface diffusion. Our main contributions are as follows:

- We formulate subclass-aware hybrid prototypes by blending dataset-specific visual prototypes with CONCH pathology text prototypes and matching them to multi-scale features from a frozen UNI encoder.
- We propose an activation-balancing operator that performs class-wise power recalibration while exactly preserving each pixel's original total foreground mass, thereby changing relative class allocation without creating or suppressing foreground evidence.
- We propose a one-step graph diffusion mechanism that regularizes feature affinity with an attention-context discrepancy from late UNI blocks; this cue is treated as a soft structural proxy rather than an explicit boundary detector.
- We distinguish direct stage-1 pseudo-mask evaluation from downstream stage-2 segmentation and evaluate the framework on BCSS-WSSS and LUAD-HistoSeg, with limitations of cross-backbone comparisons stated explicitly.

2 Methodology

Overview. We study weakly supervised semantic segmentation (WSSS) in histopathology, where each image x has only an image-level multi-label annotation $y \in \{0,1\}^C$ over parent tissue classes. Stage 1, which constitutes ProBAG, combines hybrid prototype CAM learning, activation-balanced CAM fusion, and Boundary-Aware Graph Diffusion (BAGD) to generate dense pseudo-masks without pixel-level supervision. Tables 2–4 evaluate these stage-1 masks directly. For the complete two-stage comparison in Table 1, the generated masks supervise a downstream Phikon-FPN model with LoRA; this stage-2 segmenter is an evaluation component rather than part of the proposed pseudo-mask generator. Stage-1 mask generation uses frozen UNI features and CONCH semantic guidance and does not rely on CRF or an external segmentation model.

2.1 Hybrid Prototype CAM Learning

A frozen UNI ViT-L/16 pathology foundation encoder E extracts dense ViT tokens from the input image [7,5]. We take intermediate encoder blocks and transform them with a lightweight feature pyramid [14],

$$\{F_\ell\}_{\ell=1}^{4} = \mathrm{FPN}(E(x)), \quad F_\ell \in \mathbb{R}^{B\times d_\ell \times H_\ell \times W_\ell}, \tag{1}$$

where lower levels retain finer spatial detail and the deepest level provides semantic features for graph refinement. Only the feature pyramid, prototype projections, and scale logits are trainable. The backbone roles are complementary: UNI supplies dense pathology features, CONCH supplies pathology-aligned text semantics, frozen MedCLIP is used only to encode CAM-mined regions for the

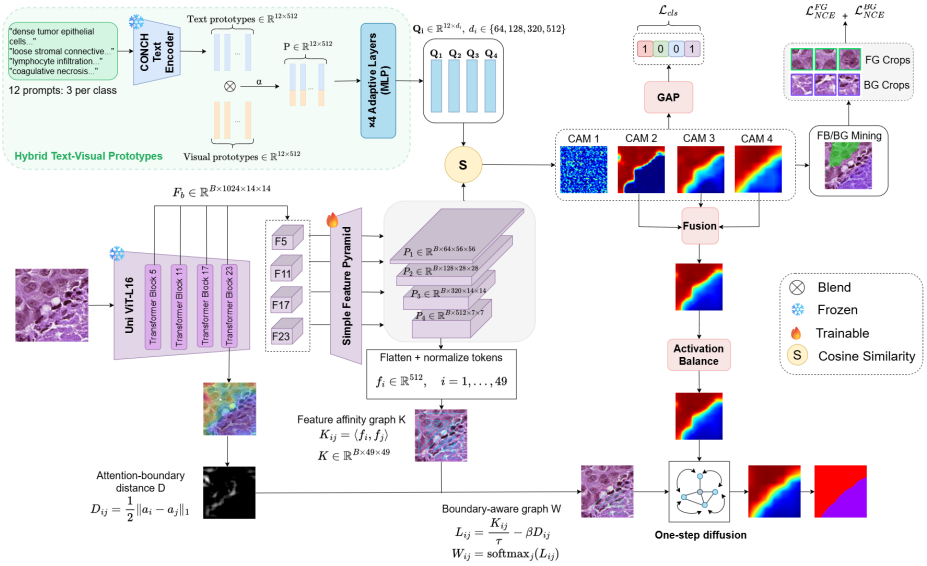

Fig. 1. Overview of the ProBAG framework. Hybrid text–visual prototypes are matched with multi-scale features from a frozen pathology encoder to generate tissue CAMs, which are then fused, activation-balanced, and refined by boundary-aware graph diffusion to produce pseudo masks.

contrastive auxiliary loss, and Phikon is reserved for the downstream stage-2 segmenter with LoRA. These models therefore occupy distinct functional roles, but the present study does not provide a controlled comparison against alternative pathology foundation models.

We represent each parent tissue class by several subclasses. Let $\mathcal{K}_c$ be the set of subclasses of parent class c, and let $\pi(k)$ map subclass k to its parent. For each subclass, a visual prototype v_k from the prototype bank is blended with a pathology text prototype t_k obtained from CONCH ViT-B/16 [15]:

$$p_k = \frac{\alpha \bar{t}_k + (1-\alpha)\bar{v}_k}{\|\alpha \bar{t}_k + (1-\alpha)\bar{v}_k\|_2}, \tag{2}$$

where $\bar{t}_k$ and $\bar{v}_k$ are ℓ_2-normalized and α controls the fixed text–visual contribution. A level-specific projection head maps the hybrid prototype to each pyramid dimension, yielding $p_{\ell,k}$.

At scale ℓ, subclass CAM logits are computed by scaled cosine similarity between normalized pixel features and projected prototypes:

$$A_{\ell,k}(u) = s_\ell \left\langle \frac{F_\ell(u)}{\|F_\ell(u)\|_2}, \frac{p_{\ell,k}}{\|p_{\ell,k}\|_2} \right\rangle. \tag{3}$$

Global average pooling gives image-level subclass logits $q_{\ell,k}$. Subclass logits are merged into parent predictions,

$$q^{\mathrm{par}}_{\ell,c} = \mathrm{Merge}(\{q_{\ell,k} : k \in \mathcal{K}_c\}), \tag{4}$$

During stage-1 classification training, Merge is the arithmetic mean over subclasses, $q^{\mathrm{par}}_{\ell,c} = |\mathcal{K}_c|^{-1} \sum_{k\in\mathcal{K}_c} q_{\ell,k}$. During pseudo-mask generation, subclass CAMs are instead aggregated by a softmax-weighted sum over subclass responses (the implementation option named "max") before parent-level normalization.

and optimized only with image-level labels:

$$\mathcal{L}_{\mathrm{cls}} = \sum_{\ell=1}^{4} \rho_\ell \, \mathrm{BCE}(q^{\mathrm{par}}_\ell, y). \tag{5}$$

CAM-guided prototype regularization. The deepest subclass CAMs are also used to mine foreground and background image regions. A dynamic threshold selects foreground regions for active subclasses; the complementary regions serve as background. These masked regions are encoded by a frozen MedCLIP vision encoder [25] and trained with foreground/background InfoNCE losses [17] against the corresponding tissue prototypes. The total objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{nce}}(\mathcal{L}_{\mathrm{FG}} + \mathcal{L}_{\mathrm{BG}} + \mu\mathcal{R}_{\mathrm{cam}}), \tag{6}$$

where $R_{\mathrm{cam}} = \mathrm{mean}(A_4)$, $\lambda_{\mathrm{nce}} = 0.05$, $\mu = 5 \times 10^{-4}$, and the InfoNCE temperature is 0.07; the dynamic CAM threshold ratio is 0.5.

2.2 Activation-Balanced CAM Fusion

For pseudo-mask generation, subclass CAMs are first merged into parent CAMs using the softmax-weighted subclass aggregation described above. Absent classes are masked using the image-level labels, and each parent CAM is independently normalized before resizing to the input resolution. We then fuse the intermediate and deep CAMs,

$$S_c = \sum_{\ell\in\mathcal{M}} \omega_\ell A^{\mathrm{par}}_{\ell,c}, \qquad \sum_{\ell\in\mathcal{M}} \omega_\ell = 1, \tag{7}$$

where $\mathcal{M}$ denotes the selected CAM scales; in our implementation $\mathcal{M} = \{2, 3, 4\}$, since the first-level CAM is spatially detailed but less semantically stable. The fused CAMs are not directly comparable across tissue classes, since some tissues produce broader or stronger activations than others. We therefore reshape class competition by class-wise power calibration while preserving the original foreground mass $m(u) = \sum_j S_j(u)$:

$$\widehat{S}_c(u) = m(u)\frac{(S_c(u)+\epsilon)^{\gamma_c}}{\sum_j (S_j(u)+\epsilon)^{\gamma_j}}, \qquad \widetilde{S}_c = (1-\eta)S_c + \eta\widehat{S}_c. \tag{8}$$

This recalibrates class competition while preserving the foreground activation mass.

2.3 Boundary-Aware Graph Diffusion

The balanced CAMs are refined by BAGD, which builds a graph from the deepest feature map F_4, following the broader use of graph-based propagation in representation learning [11]. Let f_i denote the normalized feature token at graph node i. Feature affinity is computed as

$$K_{ij} = \langle f_i, f_j \rangle, \qquad \mathcal{N}(i) = \{j : K_{ij} \geq \delta\} \cup \{i\}. \tag{9}$$

Feature similarity alone may propagate scores across tissue interfaces. To reduce such leakage, we extract late-block self-attention profiles [23] from the frozen encoder, pool them to the F4 graph resolution, and normalize them. We use the discrepancy between pooled profiles as an attention-context distance—a soft proxy for structural separation rather than an explicit boundary detector:

$$D_{ij} = \frac{1}{2}\|a_i - a_j\|_1. \tag{10}$$

A large D_{ij} indicates that two nodes participate in different global attention contexts. Subtracting βD_{ij} lowers the diffusion probability across these contextually distinct edges, while K_{ij} retains local appearance affinity. The boundary-aware diffusion graph is obtained by

$$L_{ij} = K_{ij}/\tau - \beta D_{ij}, \quad j \in \mathcal{N}(i), \qquad W_{ij} = \frac{\exp(L_{ij})}{\sum_{r\in\mathcal{N}(i)} \exp(L_{ir})}. \tag{11}$$

For each class, the balanced CAM is downsampled to graph resolution as $s_c \in \mathbb{R}^n$ and refined by one residual diffusion step:

$$z_c = (1 - \lambda_c)s_c + \lambda_c W s_c. \tag{12}$$

The interpolation weight $\lambda_c \in [0, 1]$ controls the residual smoothing strength. In the released BCSS configuration, $\lambda_c = (0.35, 0.35, 0.20, 0.45)$ for (TUM, STR, LYM, NEC), the graph resolution is 7×7, $\delta = 0.55$, $\tau = 0.10$, $\beta = 2$, and the attention profile is extracted from block 23. Refined maps are upsampled and converted to pseudo-masks by foreground argmax,

$$\hat{Y}(u) = \arg \max_{c\in\{1,\ldots,C\}} Z_c(u). \tag{13}$$

We deliberately use one residual diffusion step: the $(1 - \lambda_c)s_c$ term retains the original CAM evidence while $\lambda_c W s_c$ applies a single context-constrained correction. Repeated propagation increasingly behaves as a low-pass filter and may erase small, irregular histological structures spanning only a few graph nodes.

3 Experiments

3.1 Datasets

We evaluate on two histopathology WSSS datasets: BCSS-WSSS [3] and LUAD-HistoSeg [10]. BCSS-WSSS contains 31,826 H&E patches (train: 23,422, val:

3,418, test: 4,986) covering tumor (TUM), stroma (STR), lymphocyte (LYM), and necrosis (NEC). LUAD-HistoSeg has 17,291 patches (train: 16,678, val: 306, test: 307) covering tumor epithelium (TE), necrosis (NEC), lymphocytes (LYM), and tumor-associated stroma (TAS). Per standard WSSS, training uses only image-level labels; val/test sets use pixel masks.

3.2 Implementation Details

Stage-1 implementation. Experiments were run on an RTX PRO 4000 (24 GB). Frozen UNI blocks 5, 11, 17, and 23 [5] provide the four feature levels. Each parent class uses three visual/text subclasses, with fixed hybrid weight $\alpha = 0.6$. The released BCSS configuration uses AdamW with base learning rate 5×10^{-5}, weight decay 10^{-3}, effective batch size 16, classification scale weights (0, 0.1, 1, 1), $\lambda_{\text{nce}} = 0.05$, and validation-mIoU checkpoint selection. Pseudo-masks fuse scales 2–4 with weights (0.3, 0.3, 0.4) and use the activation-balance and BAGD settings reported above.

Stage-2 downstream evaluation. For Table 1 only, stage-1 pseudo-masks supervise a Phikon ViT-B backbone with an FPN decoder. The Phikon backbone is frozen except for LoRA adapters (rank 8, $\alpha_{\text{LoRA}} = 8$), and the model is trained for 80 epochs with batch size 16 and AdamW learning rate 6×10^{-5} using noise-robust Bootstrapped CE (noise coefficient 0.3) plus Dice loss. The best checkpoint is selected by validation mIoU. Baselines use their official downstream implementations; therefore Table 1 compares complete systems rather than isolating pseudo-mask quality under a common stage-2 architecture.

3.3 Comparison with State of the Art

Quantitative Results. Table 1 reports our retrained baseline runs using official implementations and the dataset splits/evaluation pipeline used in this study; the values are therefore not copied from the original papers and their ordering is protocol-dependent. This is particularly visible on LUAD-HistoSeg, where our retrained CAM baseline exceeds TPRO and PBIP. On BCSS-WSSS, ProBAG improves over PBIP by +4.13 mIoU, +1.52 mDice, and +4.40 FwIoU. On LUAD-HistoSeg, it improves over the strongest retrained baseline, CAM, by +2.15 mIoU, +1.38 mDice, and +1.84 FwIoU. These margins describe complete two-stage systems and may reflect both pseudo-mask quality and differences in foundation/downstream backbones; Tables 2–4 provide the more direct evidence for the proposed stage-1 components.

Qualitative Results. Figure 2 Figure 2 visually suggests that CAM, Grad-CAM, and PBIP can exhibit structural misalignment, diffuse interfaces, and pixel-level noise, whereas ProBAG often produces more coherent regions and sharper-looking transitions. Compared with TPRO, it also preserves several fine structures that are missing in the shown examples. These visualizations are consistent with the intended role of BAGD, but they are qualitative and do not substitute for a boundary-sensitive quantitative metric.

Table 1. Quantitative comparison of final segmentation performance on BCSS-WSSS and LUAD-HistoSeg. Results are from a single run for parity with baselines (ablations report mean±std). *Note:* PBIP's LUAD results (marked with †) reflect stage-1 pseudo-mask evaluation, as we could not reproduce its stage-2.

BCSS–WSSS							
Method	Metrics (%)			Per-class IoU (%)			
	mIoU	mDice	FwIoU	TUM	STR	LYM	NEC
CAM (CVPR'16) [28]	63.32	77.29	68.53	72.98	68.13	56.19	55.99
Grad-CAM (ICCV'17) [20]	58.04	72.68	66.33	71.46	66.57	53.81	40.31
Proto2Seg (IPMI'23) [18]	57.42	72.24	–	63.25	58.28	53.27	54.89
TPRO (MICCAI'23) [27]	65.50	78.90	69.42	77.40	65.10	56.20	63.40
PBIP (CVPR'25) [21]	69.03	82.84	71.93	78.13	65.56	62.11	**70.31**
Ours (ProBAG)	**73.16**	**84.36**	**76.33**	**81.87**	**75.31**	**66.29**	69.14
LUAD–HistoSeg							
Method	Metrics (%)			Per-class IoU (%)			
	mIoU	mDice	FwIoU	TE	NEC	LYM	TAS
CAM (CVPR'16) [28]	74.26	85.19	73.32	75.57	78.08	73.69	69.69
Grad-CAM (ICCV'17) [20]	71.19	83.13	71.89	75.92	68.89	72.09	67.86
Proto2Seg (IPMI'23) [18]	61.64	75.67	–	69.58	43.64	70.63	62.70
TPRO (MICCAI'23) [27]	72.90	84.27	73.14	74.30	68.80	**77.30**	**71.20**
PBIP (CVPR'25) [21]†	66.05	79.39	62.55	60.70	71.80	72.50	59.20
Ours (ProBAG)	**76.41**	**86.57**	**75.16**	**77.68**	**82.03**	74.94	71.00

3.4 Ablation Study

Effect of Main Components. Table 2 shows that pathology text semantics provide the dominant improvement: adding text prototypes raises mIoU from 68.72% to 72.90% (+4.18) and mDice from 81.22% to 84.20% (+2.98). Blending text and visual prototypes adds a smaller +0.15 mIoU and +0.10 mDice. Adding the final activation-balance and BAGD configuration raises performance by a further +0.43 mIoU and +0.28 mDice. Thus, the measured gains are complementary but not equal in magnitude; most of the improvement comes from pathology-aligned semantic prototypes, while calibration and graph refinement provide smaller corrections. The stage-1 pseudo-masks reach 73.48% mIoU, close to and slightly above the 73.16% stage-2 BCSS result. This difference may reflect pseudo-label noise and downstream optimization, but the current experiments do not establish a unique cause. Stage 1 is therefore reported as a CRF-free standalone output, while stage 2 is retained for comparison with two-stage systems.

Table 2. Effect of Main Components on initial pseudo masks for BCSS-WSSS.

Setting	mIoU	mDice
Visual baseline	68.72 ± 0.67	81.22 ± 0.49
+ Text prototype	72.90 ± 0.14	84.20 ± 0.10
+ Hybrid prototype	73.05 ± 0.09	84.30 ± 0.06
+ Hybrid + BAGD	$\mathbf{73.48 \pm 0.20}$	$\mathbf{84.58 \pm 0.14}$

Effect of Text Encoder. Within the same ProBAG stage-1 framework, Table 3 shows that domain-specific language representations matter: replacing CLIP-style text features with CONCH increases mIoU by +1.42% and mDice

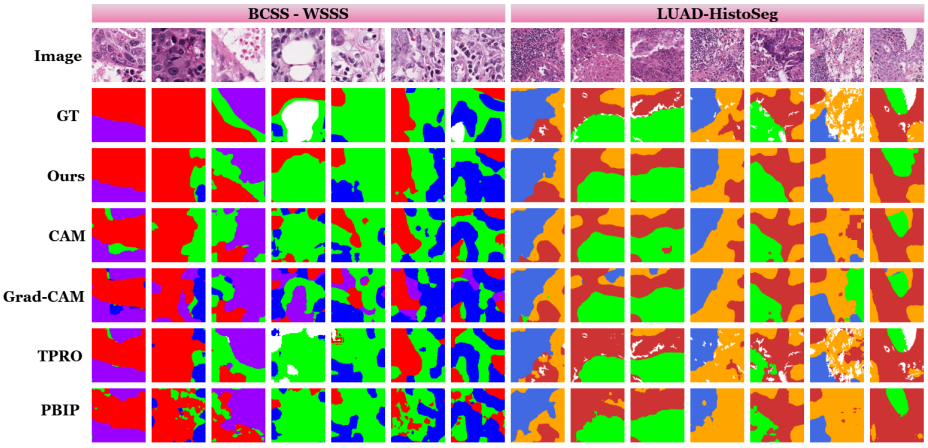

Fig. 2. Qualitative comparison of final segmentation results on BCSS-WSSS and LUAD-HistoSeg test patches. GT denotes the ground truth segmentation.

by +0.99%. This controlled comparison supports the use of pathology-aligned text semantics, while not constituting a broader comparison of visual foundation backbones.

Table 3. Effect of Text Encoders on initial pseudo masks for BCSS-WSSS.

Text Encoder	mIoU	mDice
No text	69.55 ± 0.77	81.82 ± 0.54
CLIP	72.06 ± 0.40	83.59 ± 0.27
CONCH	$\mathbf{73.48 \pm 0.20}$	$\mathbf{84.58 \pm 0.14}$

Effect of Boundary Penalty. Setting $\beta = 0$ retains feature-only graph diffusion, whereas $\beta > 0$ adds the attention-context penalty. Across $\beta \in [0, 3]$, the region-based metrics vary by less than 0.2% mIoU and 0.1% mDice. In particular, the $\beta = 0$ to $\beta = 2$ gap is 0.09% mIoU and 0.05% mDice, smaller than the reported standard deviations. These results show that the overall diffusion pipeline is stable to β, but they do not by themselves establish a statistically separable benefit of the attention penalty on region metrics. We use $\beta = 2$ as a moderate fixed constraint rather than claiming it is uniquely optimal.

Table 4. Effect of the attention-boundary penalty β on initial pseudo masks for BCSS-WSSS.

β	mIoU	mDice
0	73.39 ± 0.22	84.53 ± 0.16
1	73.43 ± 0.19	84.55 ± 0.14
2	73.48 ± 0.20	84.58 ± 0.14
3	$\mathbf{73.50 \pm 0.23}$	$\mathbf{84.60 \pm 0.15}$

3.5 Discussion and Limitations

ProBAG combines established ingredients—prototype learning, multi-scale CAMs, and graph propagation—but targets two specific histopathology failure modes with explicit mechanisms: foreground-mass-preserving class recalibration and attention-context-regularized diffusion. The current evidence is strongest for pathology-aligned text prototypes; the additional region-metric gain of the attention penalty is modest and should be complemented by boundary-specific evaluation. Table 1 compares complete systems with different foundation and stage-2 backbones and is reported from single runs, so it cannot isolate pseudo-mask quality as cleanly as the stage-1 ablations. We also do not conduct an exhaustive comparison of pathology foundation models. These limitations motivate common-backbone evaluation, multi-seed final results, and dedicated boundary metrics.

4 Conclusion

We present ProBAG, a stage-1 weakly supervised histopathology segmentation framework that combines hybrid text–visual prototypes, foreground-mass-preserving activation balance, and attention-context-regularized graph diffusion. Results on BCSS-WSSS and LUAD-HistoSeg show consistent improvements over the retrained baselines, with ablations indicating that pathology-aligned text semantics account for the largest gain and graph refinement provides a smaller complementary correction. ProBAG produces pseudo-masks without CRF or an external segmentation model; a downstream Phikon-FPN is used only for the complete two-stage comparison. Future work will evaluate boundary-sensitive metrics, common-backbone controls, multi-seed final systems, stronger foundation-model boundary cues, and large-scale multi-center data.

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations (2019), https://arxiv.org/abs/1904.05044
2. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation (2018), https://arxiv.org/abs/1803.10464
3. Amgad, M., Elfandy, H., Hussein, H., Atteya, L.A., Elsebaie, M.A.T., Abo Elnasr, L.S., Sakr, R.A., Salem, H.S.E., Ismail, A.F., Saad, A.M., Ahmed, J., Elsebaie, M.A.T., Rahman, M., Ruhban, I.A., Elgazar, N.M., Alagha, Y., Osman, M.H., Alhusseiny, A.M., Khalaf, M.M., Younes, A.A.F., Abdulkarim, A., Younes, D.M., Gadallah, A.M., Elkashash, A.M., Fala, S.Y., Zaki, B.M., Beezley, J., Chittajallu, D.R., Manthey, D., Gutman, D.A., Cooper, L.A.D.: Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics **35**(18), 3461–3467 (09 2019). https://doi.org/10.1093/bioinformatics/btz083, https://doi.org/10.1093/bioinformatics/btz083

4. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A.P., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**, 1301 – 1309 (2019), https://api.semanticscholar.org/CorpusID:196814162
5. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
6. Chen, Z., Wang, T., Wu, X., Hua, X.S., Zhang, H., Sun, Q.: Class re-activation maps for weakly-supervised semantic segmentation (2022), https://arxiv.org/abs/2203.00962
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://arxiv.org/abs/2010.11929
8. Fang, Z., Chen, Y., Wang, Y., Wang, Z., Ji, X., Zhang, Y.: Weakly-supervised semantic segmentation for histopathology images based on dataset synthesis and feature consistency constraint. In: AAAI Conference on Artificial Intelligence (2023), https://api.semanticscholar.org/CorpusID:259735474
9. Fu, M., Fu, F., Ling, X., Yuan, H., Guan, T., He, Y., Zhu, L.: Multimodal prototype alignment for semi-supervised pathology image segmentation (2025), https://arxiv.org/abs/2508.19574
10. Han, C., Pan, X., Yan, L., Lin, H., Li, B., Yao, S., Lv, S., Shi, Z., Mai, J., Lin, J., Zhao, B., Xu, Z., Wang, Z., Wang, Y., Zhang, Y., Wang, H., Zhu, C., Lin, C., Mao, L., Wu, M., Duan, L., Zhu, J., Hu, D., Fang, Z., Chen, Y., Zhang, Y., Li, Y., Zou, Y., Yu, Y., Li, X., Li, H., Cui, Y., Han, G., Xu, Y., Xu, J., Yang, H., Li, C., Liu, Z., Lu, C., Chen, X., Liang, C., Zhang, Q., Liu, Z.: Wsss4luad: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma (2022), https://arxiv.org/abs/2204.06455
11. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks (2017), https://arxiv.org/abs/1609.02907
12. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation (2016), https://arxiv.org/abs/1603.06098
13. Lee, J., Oh, S.J., Yun, S., Choe, J., Kim, E., Yoon, S.: Weakly supervised semantic segmentation using out-of-distribution data (2022), https://arxiv.org/abs/2203.03860
14. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection (2017), https://arxiv.org/abs/1612.03144
15. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024)
16. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: A survey (2020), https://arxiv.org/abs/2001.05566
17. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2019), https://arxiv.org/abs/1807.03748
18. Pan, W., Yan, J., Chen, H., Yang, J., Xu, Z., Li, X., Yao, J.: Human-machine interactive tissue prototype learning for label-efficient histopathology image segmentation (2023), https://arxiv.org/abs/2211.14491

19. Ru, L., Zhan, Y., Yu, B., Du, B.: Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers (2022), https://arxiv.org/abs/2203.02664
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**(2), 336–359 (Oct 2019). https://doi.org/10.1007/s11263-019-01228-7, http://dx.doi.org/10.1007/s11263-019-01228-7
21. Tang, Q., Fan, L., Pagnucco, M., Song, Y.: Prototype-based image prompting for weakly supervised histopathological image segmentation (2025), https://arxiv.org/abs/2503.12068
22. Van Thai, L., Nguyen, T.D., Pham, H.N., Thi, L.A.D., Nguyen, D.D., Bui, N.L.Q.: Unisemalign: Text-prototype alignment with a foundation encoder for semi-supervised histopathology segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 6803–6813 (June 2026)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023), https://arxiv.org/abs/1706.03762
24. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation (2020), https://arxiv.org/abs/2004.04581
25. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text (2022), https://arxiv.org/abs/2210.10163
26. Xie, J., Hou, X., Ye, K., Shen, L.: Cross language image matching for weakly supervised semantic segmentation (2022), https://arxiv.org/abs/2203.02668
27. Zhang, S., Zhang, J., Xie, Y., Xia, Y.: Tpro: Text-prompting-based weakly supervised histopathology tissue segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2023), https://api.semanticscholar.org/CorpusID:263673303
28. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization (2015), https://arxiv.org/abs/1512.04150