Supplementary Material

1. Evaluation with Pathology-specific Encoders

We evaluate our method using CONCH [1] as a pathology-specific patch-level feature extractor. We denote CLAM trained on CONCH features as CLAM–CONCH, and MLP trained on CONCH features as MLP–CONCH. All experimental settings (hyperparameters, training epochs, and evaluation protocols) were kept identical to those used with the ImageNet-pretrained ResNet-50. Results are shown in Table 1.

Classifier	Method	MIL-AIC ↑	MIL-SIC ↑
Random		0.371 ± 0.239	0.371 ± 0.239
CLAM (CONCH)	Gradient	0.939 ± 0.190	0.939 ± 0.173
	IG[4]	0.939 ± 0.190	0.939 ± 0.173
	IDG[5]	0.679 ± 0.339	0.679 ± 0.338
	EG[3]	0.939 ± 0.190	0.939 ± 0.173
	CIG	**0.942 ± 0.177**	**0.941 ± 0.166**
MLP (CONCH)	Gradient	0.986 ± 0.000	0.986 ± 0.002
	IG[4]	0.986 ± 0.000	0.986 ± 0.002
	IDG[2]	0.839 ± 0.243	0.840 ± 0.236
	EG[3]	0.986 ± 0.000	0.986 ± 0.002
	CIG	**0.986 ± 0.000**	**0.986 ± 0.003**

Table 1. Attribution performance on **tumor-positive** slides from the **Camelyon16** dataset using the **CONCH** encoder, evaluated with MIL-AIC and MIL-SIC metrics.

From the results in Table 1, in both CLAM-CONCH and MLP–CONCH settings, all attribution methods achieve higher MIL-AIC and MIL-SIC values, reflecting the expected effectiveness of CONCH as a feature extractor for slide-level classification. We observe that CIG performs equal to or slightly better than other attribution methods with both CLAM–CONCH and MLP–CONCH. This demonstrates that the benefits of CIG are robust to encoder choice and extend to pathology-specific feature extraction.

Figure 1 illustrates attribution maps on a representative tumor-positive slide from Camelyon16 using the CLAM–CONCH model. All IG-based methods produce visual explanations that align more closely with the ground-truth tumor segmentation compared to those generated with ResNet-50 features. In particular, CIG yields sharper and more localized highlights over tumor regions, supporting the quantitative gains reported in Table 1.

Classifier	Baseline	MIL-AIC ↑	MIL-SIC ↑
CLAM (CONCH)	Zero	0.939 ± 0.190	0.938 ± 0.175
	Mean	0.939 ± 0.190	0.938 ± 0.175
	CIG	**0.942** ± **0.177**	**0.941** ± **0.166**
CLAM (ResNet-50)	Zero	0.944 ± 0.186	0.940 ± 0.141
	Mean	0.943 ± 0.186	0.939 ± 0.143
	CIG	**0.950** ± **0.166**	**0.945** ± **0.128**

Table 2. Comparison of attribution baselines for the **CLAM** model with **CONCH** and **ResNet-50** encoders on Camelyon16 tumor-positive slides. **CIG** denotes the opposite-class baseline, **Zero** denotes the all-zero baseline, and **Mean** denotes the dataset-mean baseline. Attribution performance is evaluated using the MIL-AIC and MIL-SIC metrics.

2. Analysis of Baseline Choices

To assess the effect of different attribution baselines, we compared three strategies: (i) **CIG**, which constructs a contrastive baseline by sampling patch features from 30 slides of the opposite class; (ii) a **Zero** baseline, where all patch features are replaced with the zero vector; and (iii) a **Mean** baseline, where the baseline is computed from the dataset mean vector. For each target slide, the opposite-class sampling strategy aggregates an equal number of patches from each reference slide to form a shared baseline pool, ensuring a balanced representation.

In terms of quantitative results, sampling from the opposite class (our original CIG strategy) achieves higher MIL-AIC and MIL-SIC scores. When using the Zero vector or Dataset Mean baselines, the scores are lower; however, the qualitative visualizations show only minor differences, suggesting that the Zero vector and Dataset Mean baselines remain usable alternatives, especially when considering the computational cost of sampling.

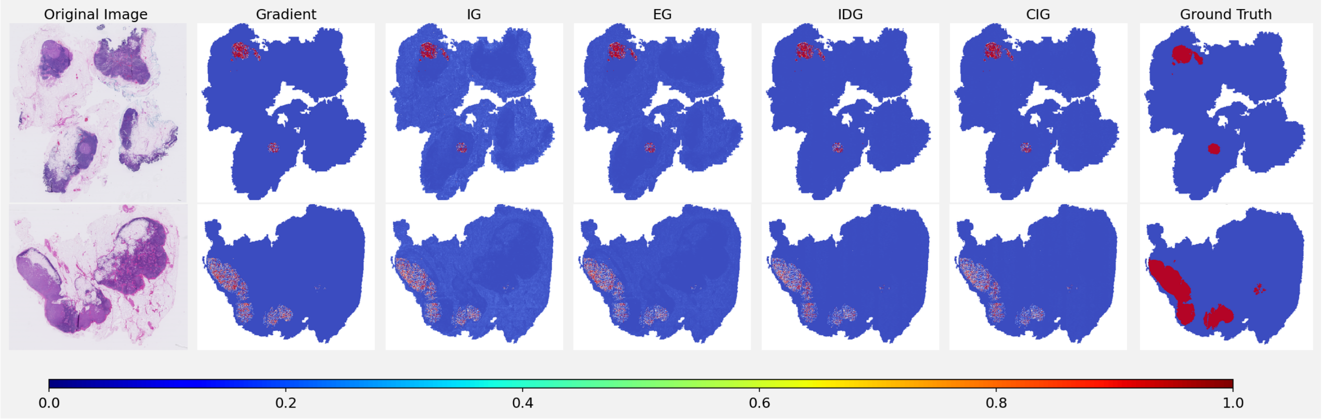

Figure 1. Qualitative comparison of attribution maps on a tumor-positive slide from the **Camelyon16** dataset using the **CLAM–CONCH** model. All Integrated Gradients (IG)-based methods show closer correspondence to the ground-truth tumor segmentation.

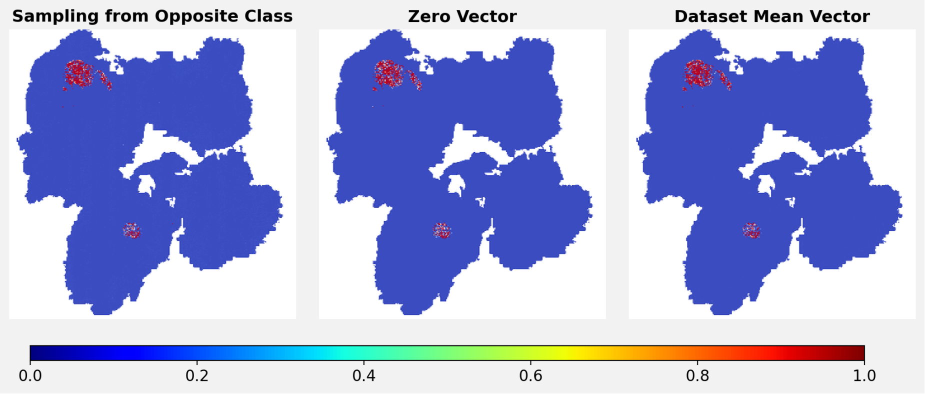

Figure 2. Visualization of a **Camelyon16** sample using the **CLAM** model with a **CONCH** encoder, comparing different baseline choices for attribution: Sampling from Opposite Class, Zero Vector, and Dataset Mean Vector.

References

[1] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024. 1

[2] Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. Integrated directional gradients: Feature interaction attribution for neural nlp models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 865–878, 2021. 1

[3] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020. https://distill.pub/2020/attribution-baselines. 1

[4] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1

[5] Chase Walker, Sumit Jha, Kenny Chen, and Rickard Ewetz. Integrated decision gradients: Compute your attributions where the model makes its decision, 2024. 1