HIS2LIP: Leveraging Weighted Inter- and Intramodal Soft Embeddings Contrastive Loss for Fine-Grained IHC Image Analysis.
—– Supplementary Material —–

Anonymous WACV Applications Track submission

Paper ID 2904

1. Datasets

Table 1 provides an overview of the datasets used for fine-tuning and testing. The fine-tuning dataset, MIHIC, contains 309,698 lung patches obtained at 40× magnification, stained with 12 biomarkers (e.g., CD3, CD20, CD34, CD38, CD68, CDK4, Cyclin-D1, D2-40, FAP, Ki67, P53, SMA) and annotated across seven tissue types (Tumor, Alveoli, Stroma, Necrosis, Immune cell, Background, Other). The test datasets span multiple organs and acquisition settings, differing substantially from MIHIC. BCData, IHC4BC, HER2-IHC-40x, and MIST focus on breast cancer, HNSCC-mIF-mIHC covers head & neck carcinoma, and PanTumor includes head & neck, lung, breast, and gastric tissues. Several datasets introduce biomarkers absent from MIHIC (e.g., ER, PR, HER2 in ACROBAT) and employ different magnifications (10×–20×). These variations in organ site, scanner type, biomarker panel, annotation protocol, and resolution create significant distribution shifts relative to MIHIC, making them strong testbeds for evaluating model robustness.

2. Caption Generation

Figure 1 illustrates the caption generation workflow. For each tissue type, representative images stained with different biomarkers were selected. An expert pathologist first provided captions for these images. Large language models (GPT-4 and Llama-3 70B) were then used to generate reformulated versions of the captions. Finally, the pathologist reviewed the augmented captions, accepting valid generations and rejecting or revising others as necessary.

3. Zeroshot Classification

Table 2 reports detailed results of zero-shot classification using both accuracy and F1-score. While accuracy is a standard metric, it can overestimate performance when class distributions are not uniform, which is common in medical datasets. F1-score, by balancing precision and recall, provides a more faithful measure of zero-shot alignment quality. For this reason, we adopt F1-score as our primary evaluation metric and use it to compare models throughout our analysis.

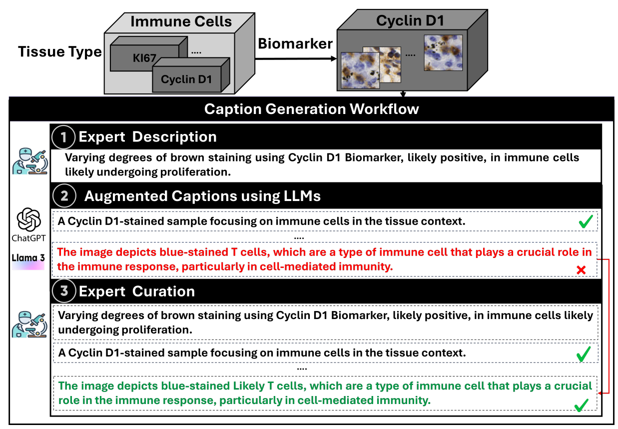

Figure 1. Caption generation workflow of the MIHIC dataset.

Not all datasets share the same classes. They are evaluated as follows:

- For **MIHIC**, **PanTumor**, and **BCData**, we perform tissue type classification as reported in Table 1.
- For **HER2-IHC**, **HNSCC-mIF-mIHC**, and **BCI**, we adopt a binary tumor vs. non-tumor classification, reflecting the positive or negative reaction of tissue to the biomarker.
- For **MIST**, **ANHIR-Lung**, and **ACROBAT**, we classify images based on unseen biomarkers. For example, in ANHIR-Lung, we classify whether an image is stained with CD31, Ki67, or ProSPC.

References

[1] Amir Akbarnejad, Nilanjan Ray, Penny J Barnes, and Gilbert Bigras. Toward accurate deep learning-based prediction of

ki67, er, pr, and her2 status from h&e-stained breast cancer images. *Applied Immunohistochemistry & Molecular Morphology*, 33(3):131–141, 2025. 3

[2] Jiří Borovec, Jan Kybic, Ignacio Arganda-Carreras, Dmitry V Sorokin, Gloria Bueno, Alexander V Khvostikov, Spyridon Bakas, Eric I-Chao Chang, Stefan Heldmann, Kimmo Kartasalo, et al. Anhir: automatic non-rigid histological image registration challenge. *IEEE transactions on medical imaging*, 39(10):3042–3052, 2020. 3

[3] P. Ghahremani, J. Marino, J. Hernandez-Prera, J. V. de la Iglesia, R. J. Slebos, C. H. Chung, and S. Nadeem. Ai-ready re-stained and co-registered multiplex dataset for head-and-neck carcinoma (hnscc-mif-mihc-comparison). The Cancer Imaging Archive, 2023. [dataset]. 3

[4] Zhongyi Huang, Yao Ding, Guoli Song, Lin Wang, Ruizhe Geng, Hongliang He, Shan Du, Xia Liu, Yonghong Tian, Yongsheng Liang, et al. Bcdata: A large-scale dataset and benchmark for cell detection and counting. In *MICCAI 2020*, pages 289–298. Springer, 2020. 3

[5] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023. 3

[6] Fangda Li, Zhiqiang Hu, Wen Chen, and Avinash Kak. Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 632–641. Springer, 2023. 3

[7] Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1815–1824, 2022. 3

[8] Md Serajun Nabi, Mohammad Faizal Ahmad Fauzi, Zaka Ur Rehman, Hezerul Bin Abdul Karim, Phaik Leng Cheah, Seow Fan Chiew, and Lai Meng Looi. Her2-ihc-40x: A high-resolution histopathology dataset for her2 ihc scoring in breast cancer. *Data in Brief*, page 111922, 2025. 3

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3

[10] Mattias Rantalainen and Johan Hartman. ACROBAT - a multi-stain breast cancer histological whole-slide-image data set from routine diagnostics for computational pathology, 2023. 3

[11] Ranran Wang, Yusong Qiu, Tong Wang, Mingkang Wang, Shan Jin, Fengyu Cong, Yong Zhang, and Hongming Xu. Mihic: a multiplex ihc histopathological image classification dataset for lung cancer immune microenvironment quantification. *Frontiers in Immunology*, 15:1334348, 2024. 3

[12] Frauke Wilm, Christian Ihling, Gábor Méhes, Luigi Terracciano, Chloé Puget, Robert Klopfleisch, Peter Schüffler, Marc Aubreville, Andreas Maier, Thomas Mrowiec, et al. Pan-tumor t-lymphocyte detection using deep neural networks: Recommendations for transfer learning in immunohistochemistry. *Journal of Pathology Informatics*, 14: 100301, 2023. 3

[13] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, pages 1–10, 2025. 3

[14] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 3

Table 1. Summary of benchmark datasets used in this study.

Dataset	Equipment	Organ	#Images	Magnification	Biomarkers	Tissue Types
MIHIC [11]	NA	Lung	309,698 patches	40x	CD3, CD20, CD34, CD38 CD68, CDK4, Cyclin-D1, D2-40 FAP, Ki67, P53, SMA	Tumor, Alveoli, Stroma Necrosis, Immune cell Background, Other
PanTumor [12]	Hamamatsu NanoZoomer	Head & Neck, Lung, Breast, Gastric	92 WSIs	40x	CD3+	Immune cell, Tumor cell, Other
BCData [4]	Motic BA600-4	Breast	1,338 patches	NA	Ki67	Tumor
HNSCC-mIF-mIHC [3]	Leica Aperio CS2	Head & Neck	1,336 patches	20x	H&E, CD3, CD8, FOXP3, PanCK	Tumor core, Tumor margin, Stroma
IHC4BC [1]	Leica Aperio GT450	Breast	90,000 patches	40x	ER, PR, Ki67, HER2	Tumor
HER2-IHC-40x [8]	3DHistech Pannoramic DESK	Breast	10,997 patches	40x	HER2	Tumor
BCI [7]	Hamamatsu NanoZoomer	Breast	4,870 patches	20x	HER2	Tumor
MIST [6]	KFBIO KF-PRO-005	Breast	22,688 patches	20x	HER2, Ki67, ER, PR	Tumor
ANHIR-Lung [2]	Zeiss Axio Imager M1	Lung	245 WSIs	10x	CD31, Ki67, ProSPC	Tumor
ACROBAT [10]	Hamamatsu NanoZoomer	Breast	4,212 WSIs	10x	H&E, ER, PGR, HER2, Ki67	Tumor

Table 2. Zero-shot classification results (Accuracy and F1-Score) across 10 datasets for SOTA VLMs and our proposed HIS2LIP models. Fine-tuned models are indicated with 🔥 and frozen models with ❄️. Best-performing results are shown in bold.

VLMs	MIHIC		BCData		PANTUMOR		HNSCC-mIF-mIHC-comparison		IHC4BC	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
CLIP [9] ❄️	30.80	25.00	33.73	18.00	28.78	13.00	26.42	21.19	53.52	45.50
PLIP [5] ❄️	18.60	11.00	38.47	16.62	28.04	13.00	30.01	23.92	51.75	38.62
BioMedCLIP [14]❄️	28.37	21.52	43.80	36.27	25.38	18.77	27.10	14.43	30.72	29.69
MUSK [13] ❄️	14.47	06.81	13.26	9.96	23.37	18.63	**35.15**	17.33	48.70	33.50
HIS2CLIP (ours) 🔥	**86.70**	**86.80**	65.93	54.89	66.97	32.52	32.78	26.46	54.37	54.35
HIS2PLIP (ours)🔥	86.41	86.50	65.53	54.51	44.97	23.77	33.68	**27.08**	**71.55**	**71.49**
HIS2BiomedCLIP (ours)🔥	84.02	84.27	**69.20**	**35.80**	**80.24**	**57.09**	31.21	19.59	61.15	58.67

VLMs	HER2-IHC-40x		BCI		MIST		ANHIR (lung)		ACROBAT	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
CLIP [9] ❄️	47.14	32.03	15.22	13.21	24.52	11.85	13.51	14.21	26.30	18.90
PLIP [5] ❄️	52.23	34.44	29.50	29.50	24.95	10.64	12.22	11.60	25.70	13.72
BioMedCLIP [14]❄️	60.45	55.28	**75.57**	44.76	13.93	11.57	16.43	16.49	**34.05**	**26.41**
MUSK [13] ❄️	59.88	50.10	17.30	33.50	25.00	10.00	**38.74**	27.05	20.10	18.00
HIS2CLIP (ours) 🔥	38.72	37.76	53.23	49.78	**27.10**	11.99	28.74	**28.35**	20.95	09.75
HIS2PLIP (ours)🔥	37.88	37.48	58.94	52.79	25.92	16.65	22.07	23.81	21.40	14.87
HIS2BiomedCLIP (ours)🔥	**64.14**	**63.80**	73.66	**56.78**	21.77	**20.13**	24.29	24.52	26.10	13.02