TC-SSA: Token Compression via Semantic Slot Aggregation for Gigapixel Pathology Reasoning

Zhuo Chen[1,2ORCID], Xiaoyu Yang[1ORCID], and Lijian Xu[1⋆ORCID]

[1] Shenzhen University of Advanced Technology, Shenzhen, Guangdong, China
xulijian@suat-sz.edu.cn
[2] University of Nottingham Ningbo China, FoSE, Ningbo, Zhejiang, China

Abstract. The application of large vision-language models to computational pathology holds great promise for diagnostic assistants but faces a critical computational bottleneck: the gigapixel scale of Whole Slide Images (WSIs). A single WSI typically contains over 10^5 patches, creating sequence lengths that exceed the constraints of standard Transformer architectures. Existing solutions often resort to spatial sampling, which risks discarding diagnostically critical evidence. To address this, we propose TC-SSA (**T**oken **C**ompression via **S**emantic **S**lot **A**ggregation), a learnable token compression framework that aggregates patch features into a fixed number of semantic slots. A gated routing module assigns patches to slots using sparse Top-2 routing, followed by weighted aggregation, enabling global slide coverage under a strict token budget. The resulting representation retains diagnostically relevant information while reducing the number of visual tokens to 1.7% of the original sequence. On SlideBench (TCGA), our model achieves 78.34% overall accuracy and 77.14% on the diagnosis subset, outperforming sampling-based baselines under comparable token budgets. The method also generalizes to MIL classification, reaching AUCs of 95.83% on TCGA-BRCA and 98.27% on TCGA-NSCLC, and an ISUP grading accuracy of 79.80% on PANDA. These results suggest that learnable semantic aggregation provides an effective tradeoff between efficiency and diagnostic performance for gigapixel pathology reasoning. Our code is available at https://github.com/OzzyChen97/TC-SSA.git.

Keywords: Gigapixel pathology · Token compression · Vision-language models.

1 Introduction

Vision-language models (VLMs) enable multimodal reasoning in computational pathology [12]. These models combine visual evidence with textual queries, offering an interface for slide-level question answering and clinical assistance. However, whole-slide images (WSIs) present a fundamental scalability challenge. A single WSI contains more than 10^5 patches, and sequence lengths exceed the memory and computational limits of standard Transformer architectures. Direct end-to-end processing of all patches is infeasible. Although recent advancements in VLMs

⋆ Corresponding author.

have enabled highly capable pathology copilots through diverse pretraining and unified architectures [9, 23, 2, 12, 13, 18, 28], deploying these comprehensive models on gigapixel slides demands a critical balance between computational efficiency and diagnostic effectiveness. To address this challenge, existing approaches predominantly follow two directions. First, spatial sampling strategies, exemplified by LLaVA-Med [11] and Quilt-LLaVA [16], restrict the input to a fixed context window by discarding the majority of patches, which introduces a severe risk of omitting diagnostically critical regions. Second, sparse-attention frameworks such as SlideChat [3] retain broader visual evidence but incur substantially higher inference costs. Parallel to these primary directions, while foundational weakly supervised modeling is frequently formulated as Multiple Instance Learning [14, 17, 5, 22], recent investigations extensively explore token compression, instruction tuning, and summarization techniques to facilitate efficient slide-level learning for pathology models [19, 4, 1, 8, 21, 24, 26]. To mitigate the efficiency bottleneck, we propose TC-SSA (Token Compression via Semantic Slot Aggregation), a learnable token-budgeting framework aggregating all patch features into fixed learnable semantic slots. To avoid systematically under-representing rare yet diagnostically critical evidence, a semantic slot regularization objective stabilizes slot utilization. Under a fixed token budget, this bottleneck maintains global slide coverage, eliminating dense attention's quadratic cost. Thus, our main contributions are: 1) **Semantic Slot-Based Token Compression:** We propose a semantic mechanism for WSI token compression, routing all visual tokens into a fixed set of learned semantic slots based on shared contextual relevance instead of spatial proximity. It aggregates sparse, diagnostically critical evidence while suppressing redundant background noise, thereby preserving the global context of the image under a rigorous token budget. 2) **Robust Regularization for Semantic Slots:** We propose semantic slot regularization to mitigate slot collapse and ensure routing stability during training. By jointly optimizing a soft load-balancing loss, an entropy regularizer, and a z-loss, the framework discourages degenerate slot flooding without forcing equal slot utilization. 3) **Superior efficiency-performance tradeoff:** Under a strict visual-token budget (only 1.7% of WSI), TC-SSA achieves 78.34% overall accuracy on SlideBench (TCGA) and improves the Diagnosis subset to 77.14%. Furthermore, it also demonstrates superior generalization with 55.94% on SlideBench (BCNB).

2 Methodology

2.1 Problem Formulation and Method Overview

We consider a WSI as a massive collection of patches. Let the visual features of these patches, extracted by the pre-trained vision encoder, be denoted as an input sequence $X \in \mathbb{R}^{B\times N\times D}$. Here, N represents the total number of patches, routinely exceeding 10^5 for a single gigapixel slide, and D denotes the feature dimension. The problem is defined as establishing a highly efficient compression function $f : \mathbb{R}^{B\times N\times D} \to \mathbb{R}^{B\times K\times D}$. This function projects the vast input sequence X into a strictly bounded representation X', where $K \ll N$ is a predefined visual token

budget. The fundamental objective is achieving this extensive token reduction while meticulously preserving the original image's global semantic context, thereby fitting dense diagnostic evidence into the restricted context window of the VLM. To resolve this bottleneck, we introduce a learnable token-budgeting framework,

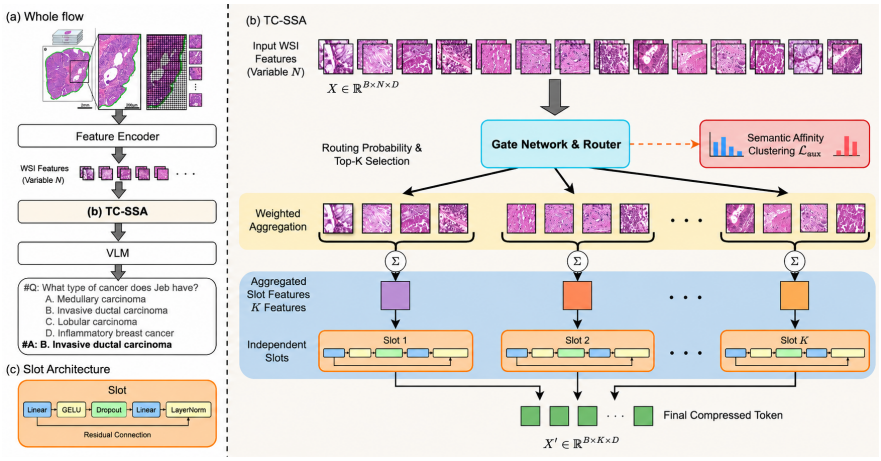

Fig. 1. Overview of the Token Compression via Semantic Slot Aggregation (TC-SSA) framework. (a) Whole flow: a feature encoder extracts a variable-length feature sequence (N) from partitioned gigapixel WSI patches. The TC-SSA module compresses these dense features into a fixed token budget before feeding them into the VLM for downstream diagnostic reasoning. **(b) TC-SSA Module:** a gate network processes input features $X \in \mathbb{R}^{B\times N\times D}$ to determine routing probabilities, followed by Top-2 routing. The auxiliary semantic slot regularization objective ($\mathcal{L}_{aux}$) stabilizes this routing distribution. Routed patch features undergo weighted aggregation into K distinct slots, yielding the final compressed representation $X' \in \mathbb{R}^{B\times K\times D}$. **(c) Slot Architecture:** the internal configuration of each independent slot.

Token Compression via Semantic Slot Aggregation (TC-SSA). As shown in Fig. 1, it comprises a gated routing mechanism and a slot-centric aggregation module. Initially, a lightweight gate computes a probability distribution over K predefined semantic slots for each input patch. To enforce sparse assignment and manage computational costs, the system applies Top-2 routing, where each patch contributes to at most two optimal slots. Subsequently, routed patches are aggregated via weighted pooling to construct compact slot embeddings. Each aggregated slot is independently refined by a multilayer perceptron to produce the final compressed sequence X'. This sequence is then projected into the downstream VLM embedding space. To maintain routing stability and prevent slot collapse during training, the framework incorporates auxiliary semantic slot regularization.

2.2 Semantic Slot Aggregation

For the j-th patch across a batch, we denote $x_j = X_{:,j,:} \in \mathbb{R}^{B\times D}$; equivalently, $x_{b,j} \in \mathbb{R}^D$ denotes the j-th patch feature of the b-th slide. A lightweight gating module computes a probability distribution over K distinct slots using shared parameters $W_g \in \mathbb{R}^{D\times K}$ and $b_g \in \mathbb{R}^K$. The resulting gate logits are $g_{b,j} = x_{b,j}W_g + b_g \in \mathbb{R}^K$, and stacking all patch logits gives $G \in \mathbb{R}^{B\times N\times K}$. The routing probability is then $P_{b,j,:} = \mathrm{Softmax}(G_{b,j,:})$. To enforce sparse assignment and control computational overhead, the system applies Top-2 routing. Rather than distributing patch information across all available slots, the framework isolates the two highest-probability slots. Let $m_{b,j,k} \in \{0, 1\}$ denote a binary mask indicating whether the k-th slot is among the Top-2 selections for the j-th patch of slide b. Consequently, truncated routing weights are defined as $\tilde{P}_{b,j,k} = m_{b,j,k}P_{b,j,k}$. This hard truncation guarantees that every patch contributes to at most two semantic concepts, thereby stabilizing the routing process and strictly controlling the subsequent aggregation cost. Following the determination of truncated routing probabilities, the framework executes a slot-centric feature aggregation step. TC-SSA does not perform explicit clustering; instead, patches with similar semantics naturally aggregate into the same slot as an emergent behavior of routing. For each slide b and semantic slot k, the aggregated feature vector $c_{b,k}$ is computed via a weighted pooling operation over all routed patches. The aggregation is formulated as

$$X'_{b,k,:} = c_{b,k} = \frac{\sum_{j=1}^{N} \tilde{P}_{b,j,k}\, x_{b,j}}{\sum_{j=1}^{N} \tilde{P}_{b,j,k} + \delta} \tag{1}$$

where $X' \in \mathbb{R}^{B\times K\times D}$ and $\delta = 10^{-9}$ is a minor numerical constant preventing division by zero if a slot receives no assignments. By normalizing the sum of the features by the sum of the routed weights, the magnitude of the resulting token remains stable regardless of the number of patches assigned to the specific slot. This mechanism effectively maps the high-dimensional visual evidence from the slide's physical coordinate space into a highly compressed, fixed-budget semantic space [25]. Through this procedure, critical diagnostic cues physically scattered across the tissue are successfully distilled into compact visual tokens without sacrificing global context.

2.3 Robust Semantic Slot Regularization

To prevent slot collapse, where a disproportionate number of patches are routed to a single semantic slot, the framework incorporates auxiliary semantic slot regularization during training. Without proper regularization, the gating module tends to heavily favor a few specific slots, which severely limits the representational capacity of the compressed sequence. To counteract this phenomenon, the primary component of the auxiliary objective is a soft load-balancing loss, denoted as

$$\mathcal{L}_{\text{switch}} = K \sum_{k=1}^{K} P_k f_k \tag{2}$$

where $P_k = \frac{1}{BN}\sum_{b=1}^{B}\sum_{j=1}^{N} P_{b,j,k}$ is the average routing probability to slot k and $f_k = \frac{1}{BN}\sum_{b=1}^{B}\sum_{j=1}^{N} m_{b,j,k}$ is the fraction of patches assigned to slot k under Top-2 routing. This loss formulation, widely adopted in sparse routing architectures, is used as a soft regularizer that penalizes degenerate cases where patches flood into one slot. It does not force equal distribution across slots: the gate can still assign background patches to a few slots while diagnostic regions occupy others. Furthermore, the auxiliary objective integrates two additional regularization terms to enhance routing stability. An entropy regularizer is applied to the gate's probability distribution to discourage overly confident but incorrect routing decisions early in training, denoted as:

$$\mathcal{L}_{\text{ent}} = 1 - \frac{-\sum_{k=1}^{K} P_k \log(P_k + \epsilon)}{\log K} \tag{3}$$

with $\epsilon = 10^{-8}$ for numerical stability. Additionally, a z-loss term $\mathcal{L}_z$ penalizes excessively large logit magnitudes produced by the gating network, thereby preventing numerical instability:

$$\mathcal{L}_z = \alpha \frac{1}{BN}\sum_{b=1}^{B}\sum_{j=1}^{N}\left(\log\sum_{k=1}^{K}\exp(g_{b,j,k})\right)^2 \tag{4}$$

where $g_{b,j,k}$ denotes the pre-softmax gate logits and $\alpha = 10^{-4}$. The auxiliary routing regularization is optimized jointly with the primary task loss, denoted as $\mathcal{L}_{task}$, of the downstream VLM. Specifically, the overall training objective is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda(\mathcal{L}_{\text{switch}} + 0.5\,\mathcal{L}_{\text{ent}} + \mathcal{L}_z) \tag{5}$$

where λ is a hyperparameter balancing the contribution of these auxiliary terms. This regularization encourages diverse and stable semantic slots, yielding compact and informative patch representations.

3 Experiments and Results

Our model is trained on SlideBench following [3], while using 2×A6000 GPUs instead of 8×A100 of [3]. The benchmark defines **Microscopy** as low-level morphology and staining, and **Diagnosis** as clinical reasoning for grading and subtyping. **Overall Accuracy** is computed across all multiple-choice VQA pairs. We use a frozen CONCH encoder [12] to extract patch features. The model uses $K = 32$ slots and auxiliary routing loss weight $\lambda = 0.1$ with Top-2 routing. **1) Efficiency-Performance Comparative Studies.** As shown in Table 1, TC-SSA achieves the best overall accuracy of **78.34%** on SlideBench (TCGA). Despite operating under the same token budget, TC-SSA surpasses SlideChat on the Diagnosis subset, highlighting the effectiveness of the routing-and-pooling strategy, which removes redundant patches while preserving critical evidence.

Table 1. Comparison of performance on SlideBench (TCGA) and zero-shot results on SlideBench (BCNB) and WSI-VQA*. SlideChat[3] is treated as the uncompressed upper bound for full WSI inference. We report FLOPs and accuracy as metrics. Best results among token-compression methods are shown in bold.

Methods	FLOPs	SlideBench (TCGA)			SlideBench	WSI
		Microscopy	Diagnosis	Overall	(BCNB)	(VQA*)
(Uncompressed Upper Bound)						
SlideChat [3]	133.3T	87.64	73.27	81.17	54.14	60.18
(Token Compression $\sim 60\times$)						
Random Baseline		24.44	24.91	25.02	24.40	24.14
GPT-4o [15]		62.89	46.69	57.91	41.43	30.41
LLaVA-Med [11]	1.70T	47.34	32.78	42.00	30.10	26.31
Quilt-LLaVA [16]	1.70T	57.76	35.96	48.07	32.19	44.43
MedDr [7]	1.70T	73.30	57.78	67.70	33.67	54.36
TC-SSA (Ours)	1.72T	**81.94**	**77.14**	**78.34**	**55.94**	**56.62**

Moreover, the superiority of TC-SSA further generalises to zero-shot settings, where it consistently outperforms competitors on both SlideBench (BCNB) and WSI-VQA*, demonstrating strong generalization of our method.

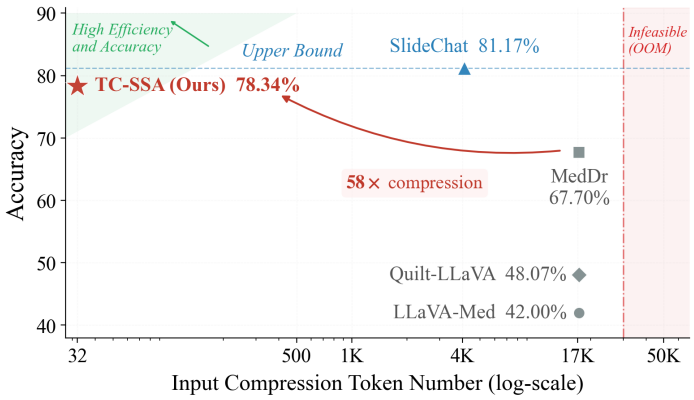

Fig. 2. Using only 32 visual tokens, TC-SSA achieves 78.34% overall accuracy on the SlideBench (TCGA) benchmark, yielding a 58× compression ratio compared to original patch features. SlideChat is reported as the uncompressed upper bound, while full-WSI inference is infeasible due to out-of-memory (OOM) constraints.

Figure 2 demonstrates superior performance on SlideBench (TCGA) of the proposed method. SlideChat serves as the uncompressed upper bound, whereas full-resolution WSI inference is impractical due to OOM constraints. By condensing global evidence into K tokens rather than discarding patches (e.g., LLaVA-Med, Quilt-LLaVA), our method outperforms sampling baselines by

10.64% in accuracy with 98.3% fewer input tokens. Its linear-time complexity $O(N \cdot K)$ enables practical deployment in memory and latency-constrained clinical settings. We further visualize slot assignments across multiple TCGA cohorts. **Figure 3** shows that each color denotes one semantic slot, and patches routed to the same slot often form coherent groups in embedding space, suggesting consistent grouping of related tissue patterns.

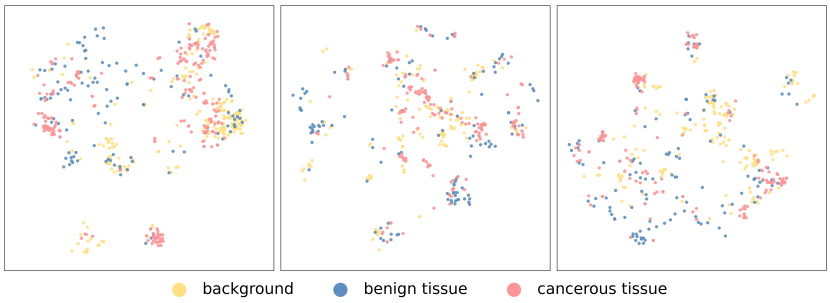

Fig. 3. Multi-dataset t-SNE visualization of semantic slot assignments. Each subplot corresponds to a different TCGA cohort. Each point is a patch embedding projected via t-SNE, with each color representing one semantic slot. The visualization shows that patches assigned to the same slot often form coherent groups, suggesting that routing captures related tissue patterns without using explicit tissue labels.

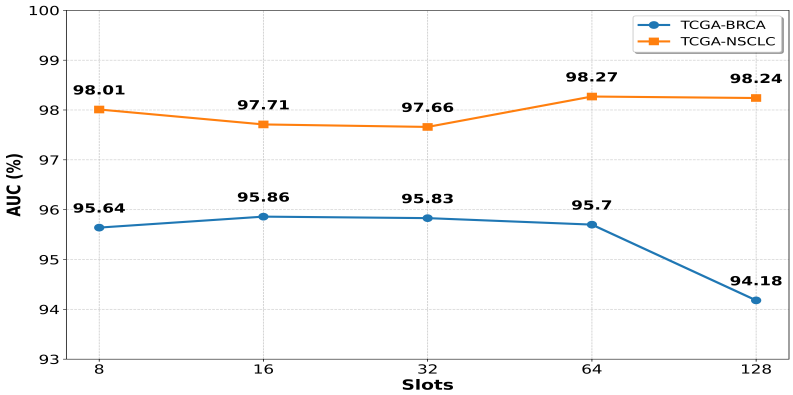

Fig. 4. Ablation experiments of the number of semantic slots K on TCGA-BRCA and TCGA-NSCLC. We set $K = 32$ by default.

2) Ablation Studies. We first ablate the number of semantic slots. Figure 4 reports MIL AUC as a function of the slot budget K. For TCGA-BRCA, per-

formance is stable for $K \in \{16, 32, 64\}$ and drops at overly large budgets (e.g., $K = 128$), suggesting diminishing returns and possible over-fragmentation of semantic evidence. For TCGA-NSCLC, larger K brings only marginal gains beyond $K = 32$. Employing Top-2 routing with an auxiliary routing loss weight $\lambda = 0.1$ stabilizes slot utilization. The auxiliary terms are used as soft semantic slot regularizers: they discourage degenerate routing collapse without forcing all slots to receive equal numbers of patches. Without sufficient $\mathcal{L}_{\text{switch}}$ regularization, reducing λ can cause inactive slots and weakened performance on heterogeneous cases. Increasing K can over-fragment semantic evidence, allowing some slots to aggregate task-irrelevant artifacts and background noise that dilute the discriminative signals passed to the final classifier. Table 2 compares ResNet-50 [6], UNI [2], and

Table 2. Encoder comparison with public MIL baselines. BRCA and NSCLC are reported by AUC, while PANDA is reported by ISUP grading accuracy. Baseline results are reported as means. TC-SSA uses the UNI encoder. Best results are shown in bold.

Encoder	Method	TCGA-BRCA	TCGA-NSCLC	PANDA
ResNet-50[6]	ABMIL[10]	83.80	92.32	58.89
	TransMIL[17]	88.52	92.49	56.42
	RRTMIL[20]	89.35	94.43	61.97
	2DMamba[27]	87.22	95.21	61.56
GigaPath[23]	ABMIL[10]	94.39	96.54	71.85
	TransMIL[17]	93.97	97.61	65.45
	GigaPath[23]	93.72	97.53	65.86
	RRTMIL[20]	94.82	97.63	72.46
	2DMamba[27]	93.84	96.87	75.72
UNI[2]	ABMIL[10]	94.05	97.04	74.69
	TransMIL[17]	93.33	97.27	68.06
	RRTMIL[20]	94.61	97.88	74.93
	2DMamba[27]	93.08	97.14	76.37
	TC-SSA (Ours)	**95.83**	**98.27**	**79.80**

GigaPath [23] encoders on TCGA-BRCA, TCGA-NSCLC, and PANDA. TCGA-BRCA and TCGA-NSCLC report cancer-subtyping AUC, while PANDA reports ISUP grading accuracy. All baseline results are reported as means, and the best results are shown in bold. Compared with the ImageNet-pretrained ResNet-50 baseline, pathology foundation encoders substantially improve performance across all datasets. Using the same UNI encoder for a controlled aggregation ablation, TC-SSA achieves the best results on TCGA-BRCA, TCGA-NSCLC, and PANDA, with gains of 1.01, 0.39, and 3.43 percentage points over the strongest baseline in each column. These results suggest that the improvement comes not only from stronger visual features, but also from the proposed semantic slot aggregation mechanism.

4 Conclusion

We presented TC-SSA, a token compression module for gigapixel pathology VLMs utilizing semantic slot aggregation. Under a strict visual-token budget, TC-SSA achieves 78.34% overall accuracy on SlideBench (TCGA), improving the *Diagnosis* subset to 77.14%, while attaining AUCs of 95.83% on TCGA-BRCA and 98.27% on TCGA-NSCLC, and 79.80% ISUP grading accuracy on PANDA. This demonstrates learnable aggregation's strong performance under constrained budgets. Currently, a fixed slide-level slot budget K makes compression quality dependent on the patch encoder, and trading fine-grained spatial geometry for semantic structure may affect localization-heavy tasks.

Acknowledgments. Zhuo Chen and Xiaoyu Yang contributed equally to this work. Zhuo Chen conducted this work during an internship at Shenzhen University of Advanced Technology. Lijian Xu is the corresponding author.

Disclosure of Interests. The authors have no competing interests to declare.

References

[1] Chen, K., Liu, M., Yan, F., Ma, L., Shi, X., Wang, L., Wang, X., Zhu, L., Wang, Z., Zhou, M., et al.: Cost-effective instruction learning for pathology vision and language analysis. Nature Computational Science (2025)

[2] Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine (2024)

[3] Chen, Y., Wang, G., Ji, Y., Li, Y., Ye, J., Li, T., Hu, M., Yu, R., Qiao, Y., He, J.: Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In: CVPR (2025)

[4] Guo, Z., Xiong, C., Ma, J., Sun, Q., Feng, L., Wang, J., Chen, H.: Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In: CVPR (2025)

[5] Hashimoto, N., Hanada, H., Miyoshi, H., Nagaishi, M., Sato, K., Hontani, H., Ohshima, K., Takeuchi, I.: Multimodal gated mixture of experts using whole slide image and flow cytometry for multiple instance learning classification of lymphoma. Journal of Pathology Informatics (2024)

[6] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)

[7] He, S., Nie, Y., Chen, Z., Cai, Z., Wang, H., Yang, S., Chen, H.: Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. CoRR (2024)

[8] Hu, Q., Lyu, W., Xu, M., Qi, K., Hu, X., Gupta, S., Zhou, J., Chen, C.: Loc-path: Learning to compress for pathology multimodal large language models. arXiv preprint arXiv:2512.05391 (2025)

[9] Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual-language foundation model for pathology image analysis using medical twitter. Nature medicine (2023)

[10] Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: ICML (2018)
[11] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In: NeurIPS (2023)
[12] Lu, M.Y., Chen, B., Williamson, D.F.K., 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 (2024)
[13] Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Zhao, M., Chow, A.K., Ikemura, K., Kim, A., Pouli, D., Patel, A., et al.: A multimodal generative ai copilot for human pathology. Nature (2024)
[14] Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering (2021)
[15] OpenAI: GPT-4o system card. arXiv preprint arXiv:2410.21276 (2024)
[16] Seyfioglu, M.S., Ikezogwo, W.O., Ghezloo, F., Krishna, R., Shapiro, L.G.: Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In: CVPR (2024)
[17] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In: NeurIPS (2021)
[18] Sun, Y., Si, Y., Zhu, C., Gong, X., Zhang, K., Chen, P., Zhang, Y., Shui, Z., Lin, T., Yang, L.: Cpath-omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. In: CVPR (2025)
[19] Tang, W., Qin, R., Fang, H., Zhou, F., Chen, H., Li, X., Cheng, M.M.: Revisiting end-to-end learning with slide-level supervision in computational pathology. In: NeurIPS (2025)
[20] Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature re-embedding: Towards foundation model-level performance in computational pathology. In: CVPR (2024)
[21] Wang, B., Zhang, K., Wang, Y., Gu, Y., Luan, H., Zhou, Y., Hu, T., Wang, R., Yang, Z., Jiang, Z., et al.: Wsisum: Wsi summarization via dual-level semantic reconstruction. Med. Image Anal. (2026)
[22] Wu, J., Chen, M., Ke, X., Xun, T., Jiang, X., Zhou, H., Shao, L., Kong, Y.: Learning heterogeneous tissues with mixture of experts for gigapixel whole slide images. In: CVPR (2025)
[23] Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature (2024)
[24] Yang, X., Xu, L., Li, H., Zhang, S.: One leaf reveals the season: Occlusion-based contrastive learning with semantic-aware views for efficient visual representation. In: ICML (2025)
[25] Yang, X., Xu, L., Zeng, X., Wang, X., Li, H., Zhang, S.: Scalar: Spatial-concept alignment for robust vision in harsh open world. Pattern Recognition (2026)

[26] Young, S., Zeng, X., Xu, L.: Fewer tokens, greater scaling: Self-adaptive visual bases for efficient and expansive representation learning. arXiv preprint arXiv:2511.19515 (2026)
[27] Zhang, J., Nguyen, A.T., Han, X., Trinh, V.Q.H., Qin, H., Samaras, D., Hosseini, M.S.: 2dmamba: Efficient state space model for image representation with applications on giga-pixel whole slide image classification. In: CVPR (2024)
[28] Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., Wang, D.: Text-guided foundation model adaptation for pathological image classification. In: MICCAI (2023)