Brain-Adapter: A Dual-Stream Vision-Language MIL Framework for Comprehensive 3D CT Diagnosis of Acute Intracranial Pathologies

Zhenyu Yi[1], Zhiyun Song[2], Yusong Sun[1], Zelin Liu[1], Manman Fei[1], Zhenhao Li[1], Jiaxuan Zhao[1], Xu Han[1], and Lichi Zhang(✉)[1]

[1] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
lichizhang@sjtu.edu.cn
[2] Department of Computing, Imperial College London, London, UK

Abstract. Automated diagnosis of 3D brain CT scans is essential for critical care, yet it remains challenging due to the heavy reliance on manual annotations and the limited semantic understanding of conventional models. While 2D foundation vision-language models (VLMs) have shown strong generalization, adapting their slice-level representations for 3D volumetric analysis remains challenging. In this paper, we propose Brain-Adapter, a novel dual-stream multiple instance learning (MIL) framework that leverages pre-trained 2D biomedical VLMs and raw diagnostic reports for robust scan-level multi-label classification. Specifically, we introduce a Text-Conditioned Attention (TCA) mechanism, utilizing raw diagnostic sentences as semantic queries to dynamically align visual cues with specific disease concepts. Concurrently, a parallel visual MIL stream captures global scan characteristics, supervised by structured labels extracted via a Large Language Model (LLM). To ensure representation coherence, a consistency constraint enforces synergy between the two streams. During inference, an Uncertainty-Aware Refinement (UAR) module dynamically calibrates and fuses these dual-stream predictions to resolve ambiguous cases. Extensive experiments demonstrate that our method significantly outperforms state-of-the-art 3D models and standard MIL approaches. By eliminating the reliance on dense annotations, Brain-Adapter offers an annotation-efficient framework for 3D acute intracranial pathology analysis. Code is available at https://github.com/Scatteredrain/Brain-Adapter.

Keywords: Vision-language model · Multiple instance learning · Acute Intracranial Pathologies.

1 Introduction

Acute intracranial pathologies encompass a highly heterogeneous spectrum of lesions, presenting as critical medical emergencies that necessitate immediate triage [18,27]. Non-contrast computed tomography (NCCT) serves as the frontline imaging modality and is routinely reconstructed into thick-slice volumes for

rapid evaluation [25]. However, the complex and diverse nature of brain lesions, coupled with the prohibitive cost of expert annotations, creates a formidable bottleneck for clinical scalability, motivating annotation-efficient and weakly supervised paradigms across medical AI [13,12,33,11,32]. Furthermore, traditional unimodal deep learning models are fundamentally constrained in this domain; they not only demand massive, exhaustively annotated datasets but also lack the broad semantic understanding required to interpret open-ended diagnostic scenarios [4,19,21].

Medical foundation vision-language models (VLMs) [28,7,30] offer a promising paradigm to circumvent manual labeling, exhibiting remarkable semantic understanding and zero-shot transferability [14,2,29]. Yet, a fundamental dimensionality gap hinders their direct application: native 3D VLMs remain rare due to the severe scarcity of paired 3D-text corpora. Multiple Instance Learning (MIL) naturally bridges this gap by treating thick-slice 3D volumes as "bags" of 2D slices [23,1,19], enabling the transfer of 2D VLM representations to 3D tasks [17]. Nevertheless, standard visual MIL approaches operate entirely within the visual domain without explicit semantic guidance. Consequently, they often struggle to pinpoint subtle pathologies, effectively attempting to localize lesions without prior knowledge of *what* specific diagnostic concepts to search for [16,3,5,8].

To address these limitations, we propose Brain-Adapter, a report-guided dual-stream MIL framework that adapts slice-level 2D VLM representations for 3D acute intracranial pathology diagnosis using raw diagnostic reports as weak supervision. We first leverage a Large Language Model (LLM) to automatically distill free-text reports into structured logic-sets. An **Open-set Alignment** branch utilizes a Text-Conditioned Attention (TCA) mechanism to retrieve pathology-relevant visual features via contrastive learning. In parallel, a **Logic-set Supervision** branch acts as a visual self-learning stream, employing an attention-based MIL [15] head supervised by the LLM-extracted labels to autonomously capture intrinsic morphological patterns across slices. A consistency constraint harmonizes these streams to ensure representation coherence, reconciling explicit semantic guidance with implicit visual evidence. During inference, an **Uncertainty-Aware Refinement (UAR)** strategy dynamically fuses the dual-stream predictions, calibrating visual outputs with text-guided evidence in ambiguous cases. We comprehensively evaluate Brain-Adapter on a real-world clinical dataset comprising complex cases. Experiments demonstrate that our method significantly outperforms state-of-the-art 3D models and standard 2D MIL approaches in multi-label classification. Moreover, to evaluate cross-domain transferability, a zero-shot abnormality detection task was conducted on the external CQ500 [4] dataset, further demonstrating the transferability of the learned disease-prompted slice aggregation under an external setting.

2 Method

We propose **Brain-Adapter**, a framework that adapts frozen 2D VLM representations to 3D volumetric analysis through report-guided slice aggregation and

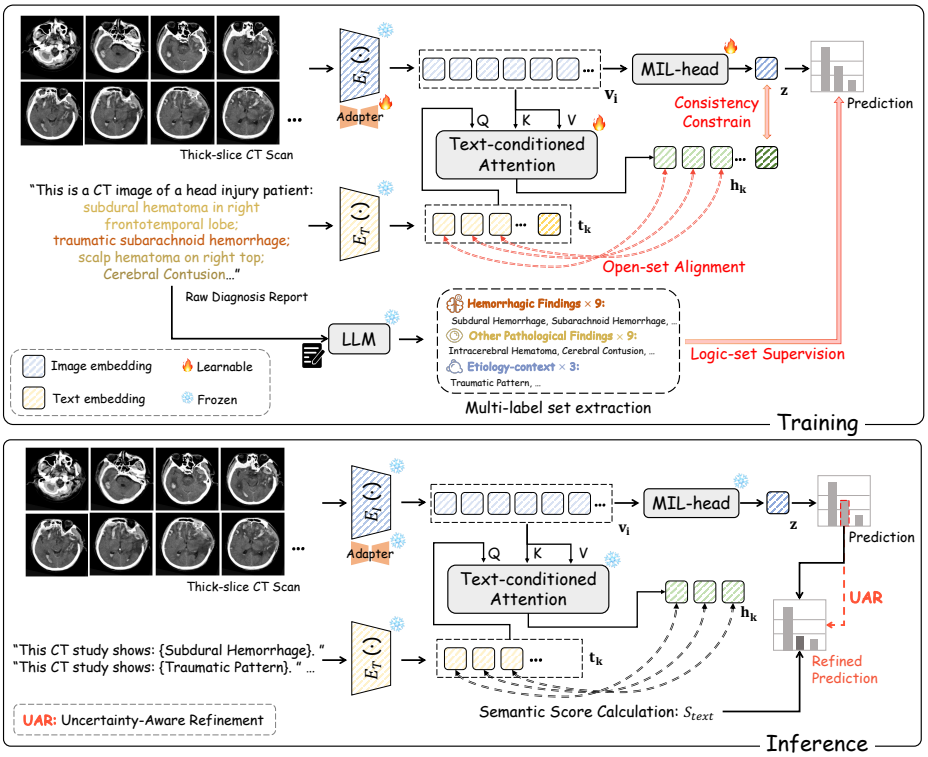

Fig. 1: Overview of Brain-Adapter. **Training (Top):** LoRA-adapted slice features feed two pathways. TCA matches text embeddings with relevant slices for *Open-set Alignment*, while ABMIL aggregates features for *Logic-set Supervision* using LLM-extracted labels. A consistency loss synchronizes global representations $\mathbf{z}$ and $\mathbf{h}_0$. **Inference (Bottom):** UAR dynamically fuses MIL predictions with prompt-based semantic scores (S_{text}) to rectify ambiguous cases.

dual-stream learning. As illustrated in Fig. 1, the architecture bifurcates into two synergistic branches during the training phase: a TCA pathway for fine-grained Open-set Alignment, and an Attention-based MIL pathway guided by Logic-set Supervision. During the inference phase, an UAR mechanism dynamically fuses the outputs of both branches to enhance diagnostic reliability.

2.1 Preliminaries: Structured Logic Extraction and Feature Encoding

Given a thick-slice CT scan $X = \{x_i\}_{i=1}^{N}$ of a patient and a raw diagnostic report R, the initial objective is to obtain structured supervision and robust visual embeddings.

LLM-driven logic-set extraction. Raw diagnostic reports R consist of the physician's final "impression", typically presenting as unstandardized, semicolon-separated diagnostic terms. We use Qwen3 to map these diagnostic conclusions to a predefined 21-label taxonomy in a strict JSON format, producing a structured multi-label set $\mathcal{Y} \in \{0,1\}^C$ ($C = 21$). The taxonomy covers Hemorrhagic Findings, Other Pathologies, and Etiology Context, forming a clinically oriented report-derived label space for the supervised branch. Detailed prompts, field definitions, and parsing examples are provided in our code repository.

Feature encoding. The contrastive pretraining stage of vision-language models enables the alignment of visual and textual latent spaces. To transfer the 2D biomedical representation capabilities to 3D CT volumes while preserving robust semantic reasoning, Low-Rank Adaptation (LoRA) [10] is applied to the self-attention blocks within the vision encoder, whereas the text encoder remains strictly frozen. The sequence of slice features is processed by the vision encoder E_V to generate visual embeddings $V = [\mathbf{v}_1, \ldots, \mathbf{v}_N] \in \mathbb{R}^{N \times d}$, where $\mathbf{v}_i = E_V(x_i)$. For the textual modality, both the complete raw diagnostic report and the subdivided fine-grained pathological descriptions are processed by the text encoder E_T. This yields global and fine-grained textual features $T = [\mathbf{t}_0, \mathbf{t}_1, \ldots, \mathbf{t}_K] \in \mathbb{R}^{(K+1) \times d}$, where $\mathbf{t}_0$ represents the embedding of the global report and $\{\mathbf{t}_k\}_{k=1}^K$ denotes the embeddings of the fine-grained sentences.

2.2 Open-set Alignment via Text-Conditioned Attention

Inspired by the clinical practice where radiologists examine specific slices to identify distinct pathologies, a **Text-Conditioned Attention (TCA)** mechanism is proposed to explicitly align visual regions with fine-grained pathological descriptions. For a specific pathological description $k \in \{1, \ldots, K\}$, the corresponding text embedding $\mathbf{t}_k$ serves as the *Query*, while the slice features V act as the *Keys* and *Values*. The pathology-specific visual representation $\mathbf{h}_k$ is aggregated as follows:

$$A_k = \text{Softmax}\left(\frac{\mathbf{t}_k V^\top}{\sqrt{d}}\right), \quad \mathbf{h}_k = A_k V \tag{1}$$

where $A_k \in \mathbb{R}^{1 \times N}$ represents the attention map indicating the relevance of each slice to pathology k.

To ensure that $\mathbf{h}_k$ accurately captures the semantics of pathology k, an **InfoNCE** loss [22] is employed for fine-grained alignment. The objective maximizes the similarity between the aggregated visual feature $\mathbf{h}_k$ and the corresponding text embedding $\mathbf{t}_k$, while distancing it from the text embeddings of other pathologies $\mathbf{t}_j$ ($j \neq k$) within the set:

$$\mathcal{L}_{align} = -\sum_{k=1}^{K} \log \frac{\exp(\cos(\mathbf{h}_k, \mathbf{t}_k)/\tau)}{\sum_{j=1}^{K} \exp(\cos(\mathbf{h}_k, \mathbf{t}_j)/\tau)} \tag{2}$$

where $\cos(\cdot)$ denotes cosine similarity and τ is a temperature parameter. This formulation forces the model to attend to slices that strictly correspond to the provided textual description.

2.3 Logic-set Supervision for Global Aggregation

To obtain a holistic scan-level representation, we employ a gated Attention-based MIL (ABMIL) [15] aggregator. Capturing complex non-linear relations, the attention weight a_i for slice i is formulated as:

$$a_i = \frac{\exp\{\mathbf{w}^\top(\tanh(\mathbf{W}\mathbf{v}_i) \odot \sigma(\mathbf{U}\mathbf{v}_i))\}}{\sum_{j=1}^{N} \exp\{\mathbf{w}^\top(\tanh(\mathbf{W}\mathbf{v}_j) \odot \sigma(\mathbf{U}\mathbf{v}_j))\}} \tag{3}$$

where $\mathbf{W}, \mathbf{U} \in \mathbb{R}^{L\times d}$ and $\mathbf{w} \in \mathbb{R}^L$ are learnable parameters, $\odot$ is element-wise multiplication, and $\sigma(\cdot)$ is the sigmoid function. The global patient representation is then aggregated as $\mathbf{z} = \sum_{i=1}^{N} a_i \mathbf{v}_i$.

Inspired by SuperCLIP [31], we explicitly supervise $\mathbf{z}$ using the structured logic-set $\mathcal{Y}$. A linear head maps $\mathbf{z}$ to class logits $\mathcal{P} = (p_1, \ldots, p_C)$. To mitigate the severe class imbalance inherent in clinical data, we optimize this multi-label classification via Asymmetric Loss (ASL) [24] instead of standard BCE. Given the predicted probability $P_c = \sigma(p_c)$, the logic-set supervision loss is:

$$\mathcal{L}_{logic} = \frac{1}{C}\sum_{c=1}^{C} [y_c\, \ell_1(P_c) + (1 - y_c)\, \ell_0(P_c)] \tag{4}$$

where $\ell_1(P_c) = -(1 - P_c)^{\lambda_1} \log(P_c)$ and $\ell_0(P_c) = -P_c^{\lambda_0} \log(1 - P_c)$. Here, $y_c \in \{0, 1\}$ is the ground truth from $\mathcal{Y}$, and λ_1, λ_0 are the focusing parameters.

Finally, to synchronize the visual and text-guided pathways, a consistency constraint aligns $\mathbf{z}$ with the global report-conditioned representation $\mathbf{h}_0$ (derived by feeding the global report embedding $\mathbf{t}_0$ into the TCA module):

$$\mathcal{L}_{cons} = 1 - \frac{\mathbf{z}^\top \mathbf{h}_0}{\|\mathbf{z}\|\|\mathbf{h}_0\|} \tag{5}$$

2.4 Uncertainty-Aware Refinement (UAR)

To rectify ambiguous MIL predictions during inference, we propose UAR to dynamically leverage the fine-grained semantic grounding of the TCA branch.

For a specific class c, let $P_c = \sigma(p_c) \in [0, 1]$ denote the predicted probability from the MIL branch. We define an uncertainty score u_c:

$$u_c = 1 - |2P_c - 1|^\alpha \tag{6}$$

where $\alpha \geq 1$ controls sensitivity, peaking ($u_c \to 1$) when the prediction is highly ambiguous ($P_c \approx 0.5$).

To compute the complementary semantic score, we insert the disease label into the prompt *"This CT study shows: {disease}."* to extract the text embedding $\mathbf{t}_c$. Using $\mathbf{t}_c$ as a query in the TCA module, we retrieve the pathology-specific visual representation $\mathbf{h}_c$. The probability-aligned semantic score is then

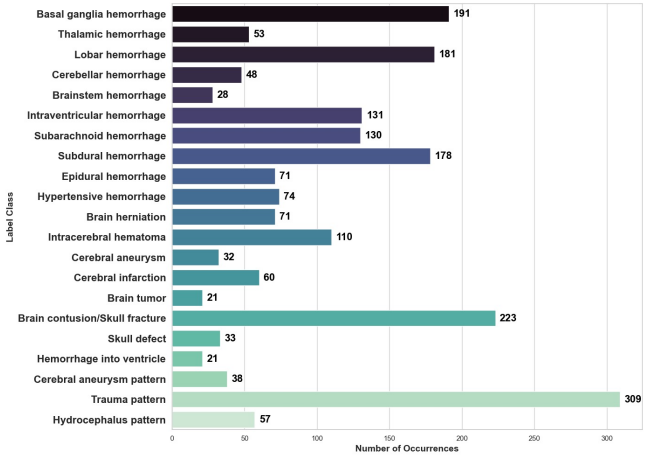

Fig. 2: Distribution of the brain pathological labels.

computed as $S_{\text{text}}^{(c)} = \sigma(\cos(\mathbf{h}_c, \mathbf{t}_c)/\tau)$. Finally, the refined prediction dynamically fuses both branches via an adaptive convex combination:

$$P_{\text{refined}}^{(c)} = (1 - \lambda \cdot u_c)P_c + (\lambda \cdot u_c)S_{\text{text}}^{(c)} \tag{7}$$

where $\lambda \in [0, 1]$ is a scaling factor. This dynamic formulation engages the text-alignment branch only when the visual prediction is uncertain. At inference, UAR uses fixed disease-label prompts rather than patient-specific diagnostic reports, so the final prediction remains CT-driven while retaining the semantic grounding learned during training.

3 Experiments and Results

Dataset and Preprocessing. We curated a real-world dataset of 852 non-contrast head CT (NCCT) studies, acquired from Siemens and GE scanners. Exclusion criteria comprised patients under 18 years of age and scans lacking axial reconstructions. A strict patient-level split allocated 682 cases for training and 170 for testing. Scans were axially aligned, resampled to $1\times1\times5$ mm, and resized to $224\times224\times32$. Hounsfield Units (HU) were clipped to $(0, 80)$ and normalized to $[0, 1]$. Training augmentations included random flipping, planar rotation ($\leq 45°$), and Gaussian noise ($\sigma = 0.01$). Additionally, Fig. 2 illustrates the prevalence distribution of the LLM-extracted pathological labels.

Implementation details. The proposed framework was built upon the pre-trained BiomedCLIP [30] architecture. LoRA adapters apply rank r = 16 with dropout probability 0.1 to the vision attention layers. Training were executed for 100 epochs on 2 NVIDIA GeForce RTX 4090 GPUs with a total batch size of 4. We use AdamW [20] as the optimizer, with a learning rate of 1e-05, a beta1 coefficient of 0.9, and a weight decay factor of 0.2.

Table 1: Multi-label classification performance on the clinical dataset. Best and second-best results are highlighted in **bold** and underlined, respectively. † denotes the variant optimized with standard BCE loss instead of ASL [24].

Method	Micro SEN	Micro SPE	Micro AUC	Macro AUC	Hamming Loss
3D Volumetric Models					
ViT-B [6]	0.567	0.780	0.759	0.534	0.240
Swin-B [9]	0.612	0.802	0.778	0.535	0.131
2D MIL Models (Trained from Scratch)					
Mean Pooling	0.266	0.977	0.775	0.493	0.094
ABMIL [15]	0.284	0.981	0.797	0.563	0.094
TransMIL [26]	0.328	0.956	0.819	0.749	0.103
2D MIL Models (BiomedCLIP [30] Pre-trained)					
Mean Pooling	0.104	**0.996**	0.840	0.661	0.087
ABMIL [15]	0.400	0.965	0.860	0.750	0.088
TransMIL [26]	0.155	0.993	0.857	0.698	0.086
Brain-Adapter†	0.285	0.982	0.876	0.773	0.084
Brain-Adapter	**0.740**	0.865	**0.887**	**0.778**	**0.079**

Comprehensive diagnosis of brain lesions. Table 1 compares Brain-Adapter against a 3D network (ViT-B [6], Swin-B [9]) and 2D MIL aggregators (Mean Pooling, ABMIL [15], TransMIL [26]) trained from scratch or initialized with BiomedCLIP [30]. Native 3D transformers like ViT-B and Swin-B struggle severely, hampered by data scarcity and the inherent difficulty of capturing fine-grained details from thick-slice CTs without explicit medical priors. Conversely, BiomedCLIP pre-training substantially boosts all MIL baselines (e.g., ABMIL Macro AUC boosts from 0.563 to 0.750), validating the efficacy of foundation models for slice feature extraction. Ultimately, Brain-Adapter achieves state-of-the-art performance (Micro/Macro AUC: 0.887/0.778, Hamming Loss: 0.079). Furthermore, compared to the BCE variant (†), employing Asymmetric Loss (ASL) prevents overwhelming true negatives from dominating gradients, drastically improving Micro Sensitivity from 0.285 to 0.740 in this highly imbalanced dataset.

Zero-shot abnormality detection. To evaluate cross-domain transferability, we apply Brain-Adapter to the external CQ500 [4] dataset (491 CT scans, ten abnormalities) without dataset-specific retraining. For each abnormality, its disease name is encoded as a textual query and used by TCA to aggregate pathology-relevant slices. As illustrated in Fig. 3(a), Brain-Adapter consistently outperforms the BiomedCLIP Mean-Pooling baseline. This result suggests that the learned text-conditioned alignment better preserves zero-shot semantic reasoning when transferring 2D VLM representations to external 3D head CT data.

Ablation study. Table 2 evaluates the contribution of each proposed module. The Open-set Alignment loss ($\mathcal{L}_{align}$) improves Micro/Macro AUC over the BiomedCLIP+ABMIL baseline by encouraging fine-grained text-visual align-

Table 2: Ablation study of our components: Open-set Alignment ($\mathcal{L}_{align}$), Consistency Constraint ($\mathcal{L}_{cons}$), and Uncertainty-Aware Refinement (UAR).

Components			Metrics		
$\mathcal{L}_{align}$	$\mathcal{L}_{cons}$	UAR	**Hamming Loss** (↓)	**Micro AUC** (↑)	**Macro AUC** (↑)
BiomedCLIP+ABMIL			0.088	0.860	0.760
✓			0.089	0.873	0.772
✓	✓		0.101	0.878	0.774
✓		✓	0.083	0.879	0.776
✓	✓	✓	**0.079**	**0.887**	**0.778**

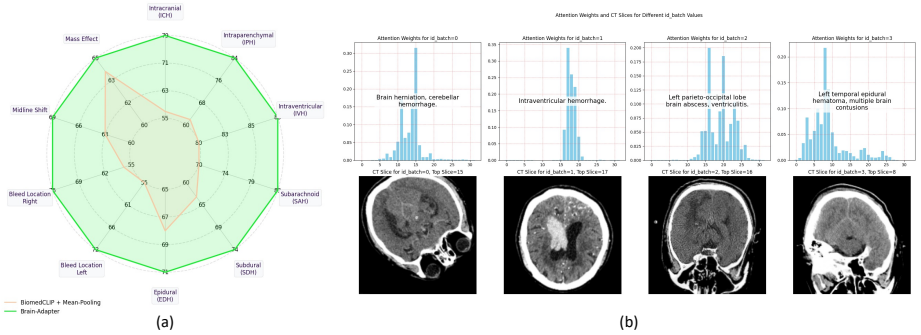

Fig. 3: Evaluation of zero-shot transferability and interpretability. **(a)** Zero-shot classification performance evaluated on the external CQ500 [4] dataset. **(b)** Explainability analysis on our clinical dataset, visualizing the text-guided slice aggregation weights derived from TCA.

ment. Adding the consistency constraint further improves AUC, although Hamming Loss increases when it is used alone, indicating a trade-off between ranking quality and thresholded multi-label errors. UAR reduces Hamming Loss by refining uncertain predictions, and combining all components yields the best overall performance.

Explainability analysis. Fig. 3(b) visualizes the TCA slice-aggregation weights as qualitative evidence of text-guided slice grounding. The attention distributions show two clinically plausible patterns. First, the model assigns higher weights to informative intermediate slices while suppressing diagnostically sparse extreme slices. Second, pathology-specific queries (e.g., *"intraventricular hemorrhage"* of the second sample) emphasize slices containing the queried finding. These observations suggest that TCA can provide slice-level evidence for model predictions, complementing its scan-level classification performance.

4 Conclusion

In this paper, we introduce Brain-Adapter, a novel dual-stream Vision-Language MIL framework that adapts 2D VLMs for the comprehensive 3D diagnosis of acute intracranial pathologies. To effectively leverage raw diagnostic reports, we construct an Open-set Alignment branch and a Logic-set Supervision branch. Furthermore, we propose an Uncertainty-Aware Refinement strategy to dynamically rectify visual predictions at test time using prompt-based semantic scores. Comprehensive experiments on a real-world clinical dataset establish a new state-of-the-art in multi-label head CT classification. Additionally, zero-shot evaluation on the external CQ500 dataset provides evidence of cross-domain transferability under disease-prompted inference. By bridging the dimensional gap between 2D foundation models and 3D clinical volumes, Brain-Adapter offers an annotation-efficient framework for acute intracranial pathology analysis.

Acknowledgments. This research is supported by the National Natural Science Foundation of China (Grant No. 62471288).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Berrimi, M., Teoh, Y.X., Chetouani, A., Houam, L., Jennane, R.: A multi-instance learning approach for improving knee osteoarthritis diagnosis from mri data. In: CBMI. pp. 1–6 (2024)
2. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
3. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., 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**(8), 1301–1309 (2019)
4. Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P., Warier, P.: Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study. The Lancet **392**(10162), 2388–2396 (2018)
5. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A.: Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine **24**(10), 1559–1567 (2018)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Eslami, S., De Melo, G., Meinel, C.: Does clip benefit visual question answering in the medical domain as much as it does in the general domain? arXiv preprint arXiv:2112.13906 (2021)

8. Fei, M., Song, Z., Shen, Z., Liu, M., Wang, Q., Zhang, L.: Uffl: Bridging slide- and cell-level annotations via unified feature fusion learning for cervical cancer screening. Information Fusion p. 104427 (2026)
9. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
11. Hu, Q., Wei, J., Yi, Z., Yan, Z., Guo, Y., Shi, H., Ji, G.P., Li, Q., Wang, Z.: Samix: Reinforcing sam2 with semantic adapter and reference selecting policy for mix-supervised segmentation. In: CVPR. pp. 17948–17958 (2026)
12. Hu, Q., Yi, Z., Zhou, Y., Huang, F., Liu, M., Li, Q., Wang, Z.: Monobox: Tightness-free box-supervised polyp segmentation using monotonicity constraint. In: AAAI. vol. 39, pp. 3572–3580 (2025)
13. Hu, Q., Yi, Z., Zhou, Y., Peng, F., Liu, M., Li, Q., Wang, Z.: Sali: Short-term alignment and long-term interaction network for colonoscopy video polyp segmentation. In: MICCAI. pp. 531–541 (2024)
14. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: ICCV. pp. 3942–3951 (2021)
15. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: ICML. pp. 2127–2136 (2018)
16. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: CVPR. pp. 14318–14328 (2021)
17. Liu, H., Georgescu, B., Zhang, Y., Yoo, Y., Baumgartner, M., Gao, R., Wang, J., Zhao, G., Gibson, E., Comaniciu, D., et al.: Revisiting 2d foundation models for scalable 3d medical image classification. arXiv preprint arXiv:2512.12887 (2025)
18. Lolli, V., Pezzullo, M., Delpierre, I., Sadeghi, N.: Mdct imaging of traumatic brain injury. The British journal of radiology **89**(1061), 20150849 (2016)
19. López-Pérez, M., Schmidt, A., Wu, Y., Molina, R., Katsaggelos, A.K.: Deep gaussian processes for multiple instance learning: Application to ct intracranial hemorrhage detection. Computer Methods And Programs In Biomedicine **219**, 106783 (2022)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
21. Neethi, A., Kannath, S.K., Kumar, A.A., Mathew, J., Rajan, J.: A comprehensive review and experimental comparison of deep learning methods for automated hemorrhage detection. Engineering Applications of Artificial Intelligence **133**, 108192 (2024)
22. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
23. Rafsani, F., Shah, J., Chong, C.D., Schwedt, T.J., Wu, T.: Dinoatten3d: Slice-level attention aggregation of dinov2 for 3d brain mri anomaly classification. In: ICCV. pp. 3544–3553 (2025)
24. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 82–91 (2021)
25. Rincon, S., Gupta, R., Ptak, T.: Imaging of head trauma. Handbook of clinical neurology **135**, 447–477 (2016)

26. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
27. Sharp, A.L., Nagaraj, G., Rippberger, E.J., Shen, E., Swap, C.J., Silver, M.A., McCormick, T., Vinson, D.R., Hoffman, J.R.: Computed tomography use for adults with head injury: describing likely avoidable emergency department imaging based on the canadian ct head rule. Academic Emergency Medicine **24**(1), 22–30 (2017)
28. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: EMNLP. pp. 3876–3887 (2022)
29. Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., et al.: A vision–language foundation model for precision oncology. Nature **638**(8051), 769–778 (2025)
30. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
31. Zhao, W., Huang, Z., Feng, J., Wang, X.: Superclip: Clip with simple classification supervision. arXiv preprint arXiv:2512.14480 (2025)
32. Zhao, X., Wang, S., Song, Z., Shen, Z., Yao, L., Yuan, H., Wang, Q., Zhang, L.: Adler: adversarial training with label error rectification for one-shot medical image segmentation. Expert Systems with Applications p. 131224 (2026)
33. Zhao, X., Li, Z., Luo, X., Li, P., Huang, P., Zhu, J., Liu, Y., Zhu, J., Yang, M., Chang, S., et al.: Ultrasound nodule segmentation using asymmetric learning with simple clinical annotation. IEEE Transactions on Circuits and Systems for Video Technology **34**(10), 9010–9023 (2024)