Towards Scalable Language-Image Pre-training for 3D Medical Imaging

Chenhui Zhao   Yiwei Lyu   Asadur Chowdury   Edward Harake
Akhil Kondepudi   Akshay Rao   Xinhai Hou   Honglak Lee   Todd Hollon
University of Michigan

Abstract

Language-image pre-training has demonstrated strong performance in 2D medical imaging, but its success in 3D modalities such as CT and MRI remains limited due to the high computational demands of volumetric data, which pose a significant barrier to training on large-scale, uncurated clinical studies. In this study, we introduce *Hierarchical attention for Language-Image Pre-training* (HLIP), a scalable pre-training framework for 3D medical imaging. HLIP adopts a lightweight hierarchical attention mechanism inspired by the natural hierarchy of radiology data: slice, scan, and study. This mechanism exhibits strong generalizability, *e.g.*, +4.3% macro AUC on the Rad-ChestCT benchmark when pre-trained on CT-RATE. Moreover, the computational efficiency of HLIP enables direct training on uncurated datasets. Trained on 220K patients with 3.13 million scans for brain MRI and 240K patients with 1.44 million scans for head CT, HLIP achieves state-of-the-art performance, *e.g.*, +32.4% balanced ACC on the proposed publicly available brain MRI benchmark Pub-Brain-5; +1.4% and +6.9% macro AUC on head CT benchmarks RSNA and CQ500, respectively. These results demonstrate that, with HLIP, directly pre-training on uncurated clinical datasets is a scalable and effective direction for language-image pre-training in 3D medical imaging. The code is available at `https://github.com/Zch0414/hlip`

1 Introduction

Language-supervised pre-training is well-suited for radiology, where each study comprises medical images paired with a corresponding radiologist's report. This natural alignment between visual and textual information has motivated the adaptation of language-image pre-training methods, such as CLIP [1, 2], to learn clinically meaningful radiology representations. CLIP-based models are especially notable for their clinical utility: they demonstrate strong zero-shot transfer performance on diagnostic tasks [3, 4], and their encoders consistently improve performance on multimodal learning benchmarks [5, 6]. In the domain of 2D medical imaging, such as chest X-rays, language-supervised pre-training is a key driver for integrating computer vision into clinical workflows [7, 8, 9].

However, language-supervised pre-training in 3D medical imaging has yet to reach the scale or performance demonstrated in 2D modalities. For instance, chest X-ray CLIP models have been trained on over 500K 2D images [8, 9, 10], achieving human-level performance on multiple diagnostic tasks [9]. In contrast, progress in 3D medical imaging remains limited, with existing models underperforming relative to their 2D counterparts [3, 5]. We attribute this performance gap between 2D and 3D language-image pre-training to three primary factors: the need for data curation and annotation, limitations in model architectures, and the complexity of 3D medical imaging studies.

Computed tomography (CT) and magnetic resonance imaging (MRI) generate 3D volumetric images across various anatomical regions, including the brain, chest, and abdomen. Unlike 2D imaging, a single clinical CT or MRI study may comprise multiple 3D volumes, often called scans or sequences, each capturing different imaging protocols or orientations. As illustrated in Figure 1, a standard
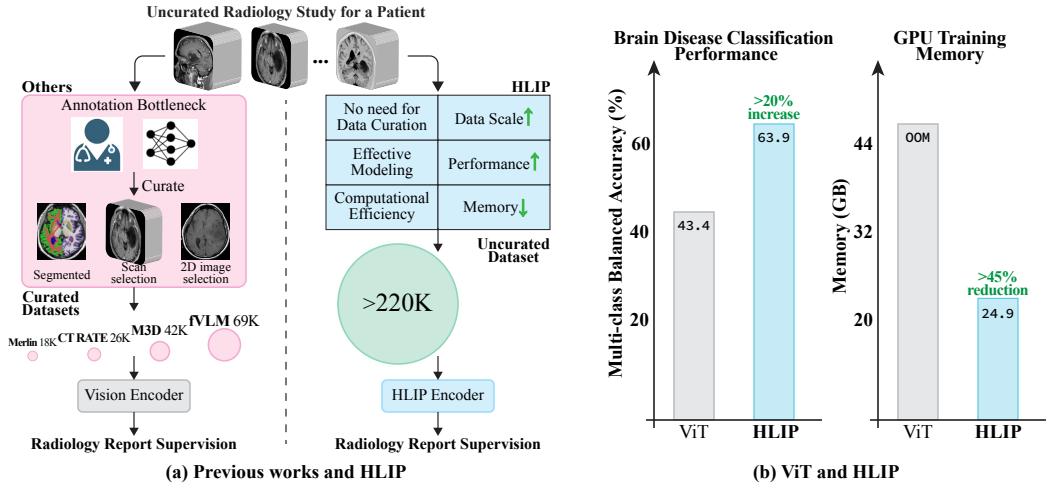
| (a) Previous works and HLIP | (b) ViT and HLIP |

Figure 1: Illustration of (**a**) an uncurated study for a patient. While previous work has relied on annotation and curation, HLIP enables language-image pre-training directly on such data. (**b**) HLIP achieves higher performance in zero-shot brain MRI disease classification while using less memory than the original ViT, which can only be trained with gradient checkpointing.

MRI study typically includes several sequences (e.g., T1-weighted, T2-weighted, and FLAIR), each contributing distinct diagnostic information. Similarly, CT studies often include scans acquired with varying orientations or scanner settings within the same study. Naively encoding such uncurated studies with standard visual encoders, such as the Vision Transformer (ViT) [11], can produce on the order of 10^4 tokens per study, resulting in substantial computational overhead. The sheer size and structural complexity of uncurated studies therefore present a significant barrier to scalable language-image pre-training. A common strategy is to curate datasets by having radiologists manually select a representative scan or 2D slice from each study, as shown in Figure 1 (a) [3, 4, 5, 6]. Alternatively, specialized model architectures, which depend heavily on preprocessing, have been proposed to reduce complexity[3]. Yet, these bespoke designs tend to generalize and scale poorly in real-world settings and are incompatible with established ViT training protocols [2, 12, 13]

In this work, we address the key barriers to scaling language-image pre-training for 3D medical imaging by directly leveraging uncurated studies and the original ViT. To effectively extract features from uncurated studies, we introduce _Hierarchical attention for Language-Image Pre-training_ (HLIP), a lightweight adaptation that leverages the natural hierarchy of radiology data: slice, scan, and study. Unlike architectural designs such as Swin [14], MViT [15, 16], and Hiera [17], HLIP utilizes the radiology data structure to define the attention scope. Furthermore, compared to efficiency-oriented modules such as window attention [18], HLIP does not trade accuracy for reduced compute. As illustrated in Figure 1 (b), our method achieves higher performance while using less memory than ViT. Moreover, HLIP is orthogonal to flash attention [19] and patch dropout [20], further alleviating the computational burden associated with uncurated studies that contain multiple 3D scans.

When trained on the public CT-RATE [3] dataset, HLIP demonstrates strong generalizability, outperforming the state-of-the-art model [21] by 4.3% macro AUC on the external Rad-ChestCT [22] dataset. Furthermore, HLIP enables scalable and effective vision-language pre-training on health system-scale datasets. Specifically, HLIP trained on our health system outperforms 2D foundation models [23, 24] on the proposed publicly available brain MRI benchmark, Pub-Brain-5, by 32.4% balanced ACC; and surpasses the head CT foundation model [25] by 1.4% and 6.9% macro AUC on the RSNA [26] and CQ500 [27] benchmarks, respectively. Our key contributions lie in three-fold:

- We introduce HLIP, a hierarchical attention mechanism that leverages the natural structure of radiology data to enable effective and scalable language-image pre-training for 3D medical imaging.

- We conduct the largest-scale training for 3D medical imaging to date, using 220K patients with 3.13 million scans for brain MRI and 240K patients with 1.44 million scans for head CT.

- We demonstrate state-of-the-art performance on multiple benchmarks spanning diverse modalities and anatomical regions, including chest CT, head CT, and brain MRI.

2

2 Related Work

Language-Image Pre-training in Radiology has been developed for 2D and 3D medical imaging, facilitated by public datasets such as CheXpert [28], MIMIC-CXR [29], and CT-RATE [3]. In 2D imaging, techniques such as local alignment [30, 31, 32], knowledge enhancement [8, 33], and longitudinal analysis [34] have been explored. CheXZero [9] demonstrates strong empirical results comparable to those of human experts in the domain of chest X-rays. BiomedCLIP [23] and ConceptCLIP [24] have been trained on more than *15 million* sample pairs, achieving strong performance across modalities, including radiology. In addition, other works [7, 10, 35] have also contributed significant insights and achieved impressive performance.

3D medical imaging offers a more comprehensive view of anatomical structures. However, due to its computational cost and limited data availability, several studies [36, 37, 38] focus on bridging the domain gap between 2D and 3D. For example, UniMedI [38] learns a shared feature space for both 2D and 3D modalities and demonstrates improvements across both. BIUD [36] distills 3D representations from a well-trained 2D model [9], significantly improving data efficiency.

More recently, aided by advances in hardware and the availability of public datasets [3], several studies [3, 5, 6, 39] have performed language-image pre-training on 3D imaging, surpassing methods [36] that still rely on 2D representations. Specifically, CT-CLIP [3] has been pre-trained on 47,149 paired chest CT scans and reports. M3D [5] explores a versatile multi-modal large language model designed for universal 3D medical imaging analysis. fVLM [6] proposes a fine-grained pre-training framework that relies on segmentation to perform organ alignment. However, while these methods have demonstrated promising performance on various tasks such as zero-shot abnormality detection and report generation, several factors limit their training scalability and real-world applicability. For example, CT-CLIP and M3D are constrained to radiology studies containing only a single imaging scan. Such studies rely heavily on human annotation and also fail to reflect real-world scenarios, where studies typically include multiple scans. fVLM further depends on a segmentation model, which introduces bias from segmentation quality and largely limits its scalability and generalizability. Moreover, these methods suffer from inefficient modeling of 3D medical imaging, leading to small batch sizes (*e.g.*, *48*), which are insufficient for effective language-image pre-training [2].

Efficient Language-Image Pre-training is crucial, as it directly affects the number of sample pairs seen during training and contrasted per batch, two key factors of model capacity. Merlin [4] and CT-CLIP [3] adopt hierarchical architectures; however, as analyzed in Appendix E.2, these models [3, 14] may be less efficient than commonly assumed when compared to the original ViT. From the architecture perspective, components such as relative position embedding are expensive for 3D inputs and is not compatible with recent advancements like flash attention [19]. From the training perspective, FLIP [20] demonstrates a favorable trade-off by randomly removing 50% of tokens during training. However, models that require a fixed activation shape [14, 16, 17, 18] cannot benefit from this strategy. Moreover, many works in radiology [6, 9, 40, 41] have demonstrated the benefits of the universal feature learned by MAE [12]. Therefore, exploring an effective approach to modeling 3D medical imaging with the original ViT [11] is a promising research direction and stands to benefit from recent advancements such as flash attention, patch dropout, and pre-trained models.

3 Method

Our goal is to perform language-image pre-training on uncurated studies using the original ViT [11]. Given a study $S \in \mathbb{R}^{M \times 1 \times D \times H \times W}$, where M denotes the number of single-channel 3D scans, and D, H, and W represent the depth, height, and width dimensions of each scan, respectively. Following the ViT [11], each scan is divided into a grid of non-overlapping volumes of size $(\frac{D}{d}, \frac{H}{h}, \frac{W}{w})$. These volumes are projected into visual tokens $F_s \in \mathbb{R}^{N \times c}$, where $N = M \times d \times h \times w$ is the total number of tokens and c denotes the number of channels. As N can be on the order of 10^4, computing self-attention over all tokens throughout the ViT backbone is prohibitive in memory. We introduce a hierarchical attention mechanism guided by the inherent data hierarchy of radiology studies, enabling more effective self-attention while preserving the ViT architecture. The strength of our approach lies in enabling scalable language-image pre-training on uncurated studies while leveraging the established training protocol of ViT [2, 12, 13].
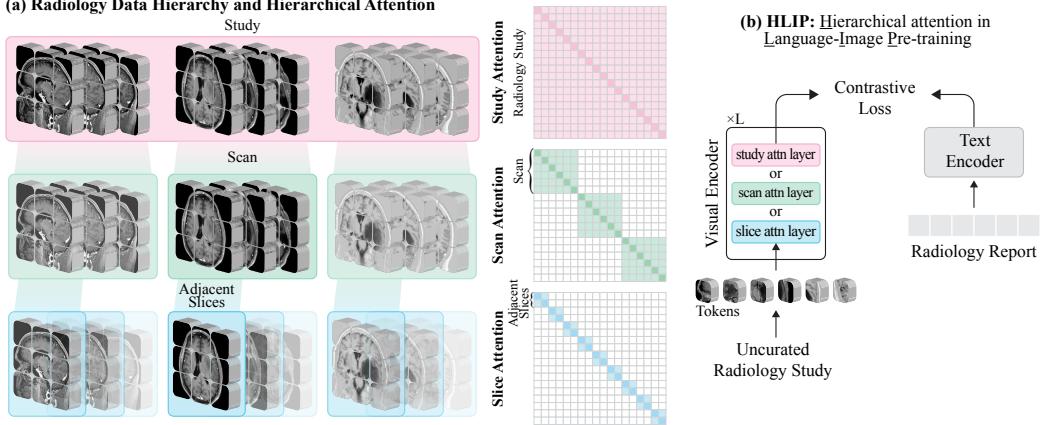
(a) Radiology Data Hierarchy and Hierarchical Attention

(b) HLIP: Hierarchical attention in Language-Image Pre-training

Figure 2: Illustration of **(a)** the radiology data hierarchy for a single patient, including the study, single scan, and adjacent slices. Our hierarchical attention mechanism mirrors this hierarchy and computes self-attention independently within each level. **(b)** Our HLIP framework incorporates a visual encoder that performs attention at different levels. In practice, lightweight slice or scan attention with a few study attention layers suffices to extract features from the full radiology study.

3.1 Hierarchical Attention Mechanism

The radiology study exhibits an inherent hierarchical structure for each patient, and our proposed hierarchical attention mechanism mirrors this structure. As illustrated in Figure 2, the study for an individual patient comprises three levels:

- The **study** contains M imaging scans, whose modalities and acquisition planes are selected by a radiologist based on the clinical context. Collectively, these M scans contain all the visual information required for radiologic diagnosis.

- The **scan** contains D slices. While a single scan conveys only partial context of the full radiology study, it still captures the complete extent of the target pathology.

- The **adjacent slices** contains $\frac{D}{d}$ consecutive slices within a single scan. Although it conveys only partial information about the target pathology, the slices can capture focal diagnostic features.

We explore a simple hierarchical attention mechanism grounded in this structure. Specifically, as shown in Figure 2, we compute self-attention independently within each hierarchical level:

- **Study attention** computes a single self-attention operation over all N tokens in a study. The I/O complexity [42] of the study attention is $\Omega(N^2 + N \times c)$ [19], as detailed in Appendix E.

- **Scan attention** computes M independent self-attention operations, each over $d \times h \times w$ tokens within a single scan. The I/O complexity of the scan attention is $\Omega(\frac{N^2}{M} + N \times c)$.

- **Slice attention** computes $M \times d$ independent self-attention operations, each over $h \times w$ tokens within a group of adjacent slices. The I/O complexity of the slice attention is $\Omega(\frac{N^2}{M \times d} + N \times c)$.

This contrasts with existing methods that modify attention computation using convolution [15, 16], pooling [17], or shifted windows [14], all of which require either regular activation shapes or attention masks. Our mechanism relies solely on a simple reshape operation, as there is no overlap between different scans or groups of adjacent slices[1], allowing greater flexibility for radiology studies that comprise multiple scans. As shown in Figure 2, each layer in the ViT can perform attention at any level. In practice, we evenly divide the ViT backbone into four subsets of layers (e.g., three layers per subset for the 12-layer ViT-B) and apply study attention only to the last layer of each subset, while the remaining layers perform the lighter scan or slice attention.

HLIP integrates seamlessly with recent efficiency advances like patch dropout [20] and flash attention [19]. Furthermore, unlike window attention [18], our scan and slice attentions do *not trade performance for efficiency* as they preserve all diagnostic features.

---

[1]With the original attention layer unchanged, study attention takes the input of size $(B, M \times d \times h \times w, c)$; scan attention takes the input of size $(B \times M, d \times h \times w, c)$; and slice attention takes the input of size $(B \times M \times d, h \times w, c)$, where B denotes the batch size.

3.2 Implementation

Model. Our visual encoder is a MAE pre-trained ViT-B [12]. The input is rescaled to [0,1] and normalized with MAE's averaged *mean* and *std*. Compared to the original ViT designed for 2D RGB images, our vision encoder only has three differences: (1) each scan is divided into 3D volumes instead of 2D patches; (2) the positional embedding encodes 3D spatial coordinates within each scan, along with an additional dimension to distinguish different scans; The *cls token* is minimally adapted to propagate information across layers that perform different hierarchies of attention.

For (1), which employs a convolution layer [12], we adopt average inflation initialization [43]. Specifically, we first sum the 2D weights along the channel dimension to accommodate a single-channel input, then replicate them $\frac{D}{d}$ times to construct the 3D weights, and finally scale by a factor of $\frac{d}{D}$. We set the token size as (8,16,16) by default. For (2), we *tile* the pre-trained 2D positional embedding $P_{mae} \in \mathbb{R}^{h \times w \times c}$ with a 1D sinusoidal positional embedding for slices, $P_{slice} \in \mathbb{R}^{d \times c}$, and another for scans, $P_{scan} \in \mathbb{R}^{M_{max} \times c}$. M_{max} indicates the maximum number of scans our model can process without interpolation. During training, to construct a batch, we randomly sample M scans along with M positional embeddings from P_{scan}, which forms the final positional embedding $P \in \mathbb{R}^{M \times d \times h \times w \times c}$. We set M_max=40 and M=10 by default. For (3), the *cls token* may lose gradient continuity across different hierarchies. To address this issue while maintaining the efficiency of our architecture, we propagate the *cls token* using a combination of cloning and averaging. For example, when transitioning from the study to scan attention layer, we distribute the *cls token* across M scans by cloning. In the reverse direction, we aggregate information from M *cls tokens* into a single *cls token* by averaging them. The transition between other hierarchies (*e.g.*, study and slice or scan and slice) follows the same procedure. Empirically, we find this strategy to be sufficient, performing better than alternatives such as weighted average or global pooling at the end of the encoder.

In our scenario, a curated dataset with only one imaging scan per study corresponds to the special case of M=1. For curated datasets, we employ slice attention with *4 evenly* distributed scan attentions; for uncurated datasets, we use scan attention with *4 evenly* distributed study attentions. We use PubMedBERT [44] as our text encoder. Model configurations are provided in Appendix B.

Datasets. Given that HLIP enables pre-training on real-world, clinical studies, we collect two datasets within our health system, namely *BrainMRI220K* and *HeadCT240K*. BrainMRI220K contains 220,993 MRI studies. We hold out 992 studies for hyperparameter tuning using the retrieval task. In the training split, the number of scans per study ranges from 1 to 162, with the third quartile at 17; the number of slices per scan ranges from 5 to 500, with the third quartile at 80. HeadCT240K contains 244,253 CT studies for training and 998 held-out studies. The number of scans per study ranges fro 1 to 71, with the third quartile at 6; the number of slices per scan ranges from 5 to 500, with the third quartile at 110. More details about these two datasets are provided in Appendix A.

Preprocessing. We do not standardize the orientation or spacing. Instead, we consistently align the depth dimension with the through-plane axis of each scan, and then resize it to a fixed shape of (48,224,224). As this differs from prior practice [3, 4, 6], we provide a discussion in Appendix A.3. For brain MRI, we apply [0.5,99.5] percentile clipping to the intensity values. For head CT, we expand each scan with Hounsfield Unit (HU) values truncated to [0,120] for soft tissue, [-20,180] for contrast-enhanced tissue and blood vessels, and [-800,2000] for bone.

Pre-training. Our implementation builds upon OpenCLIP [13] and FLIP [20]. We apply 25% patch dropout by default as regularization and acceleration. We do not apply additional augmentation or unmasked fine-tuning. Training 20 epochs on our uncurated datasets takes ~1 day with a batch size of 256 on 8 L40 GPUs. More details are provided in Appendix B.2.

4 Experiments

We apply HLIP to chest CT, brain MRI, and head CT. For chest CT and head CT, we follow prescribed evaluation protocols from previous work [3, 6, 25]. For brain MRI, we construct a new benchmark for zero-shot multi-class classification, using public datasets [21, 45, 46, 47, 48, 49, 50, 51]. To isolate the effectiveness of our hierarchical attention mechanism, we first apply HLIP to curated chest CT datasets [3, 22]. We then apply HLIP to our uncurated brain MRI and head CT datasets, highlighting the equivalent importance of scaling, modeling, and efficiency. Finally, we conduct comprehensive ablation and analysis to investigate our designs.

Table 1: Results of multi-label zero-shot evaluation on the CT-RATE [3] test split for internal validation and the Rad-ChestCT [22] for external validation. The best results are highlighted in **bold**.

Method	Internal Validation (CT-RATE)					External Validation (Rad-ChestCT)				
	AUC	ACC	F1	Precision	Recall	AUC	ACC	F1	Precision	Recall
Supervised by Original Reports										
CT-CLIP [3]	73.3	66.9	70.8	32.6	- -	63.3	59.9	64.7	34.1	- -
BIUD [36]	71.3	68.1	71.6	33.8	67.3	62.9	60.6	65.2	33.7	59.6
Merlin [4]	72.8	67.2	70.9	33.7	70.1	64.4	61.9	66.3	34.8	61.0
HLIP (ours)	**77.7**	**71.4**	**74.7**	**37.9**	**73.0**	**72.3**	**68.4**	**72.1**	**40.4**	**66.7**
Supervised by Qwen [52] Summarized Reports										
fVLM [6]	77.8	71.8	75.1	37.9	72.8	68.0	64.7	68.8	37.4	64.6
HLIP (ours)	**78.7**	**72.4**	**75.5**	**38.4**	**74.1**	**71.7**	**67.7**	**71.4**	**39.8**	**66.9**

4.1 Chest CT

Implementation Details. Following CT-CLIP [3] and fVLM [6], we apply HLIP to the CT-RATE [3] training set, then perform internal validation on its test split and external validation on the full Rad-ChestCT dataset [22]. The implementation details differ from Section 3.2 as these datasets are curated. During preprocessing, each scan is standardized with a spacing of $(3mm, 1mm, 1mm)$ [6]. The HU values are truncated to $[-1150, 350]$ [36]. We apply a center crop of size $(112, 336, 336)$ to construct mini-batch and a token size of $(8, 24, 24)$. We use CXR-BERT [7] as the text encoder. HLIP is then trained for 20 epochs without patch dropout, which can be completed within 6 hours using a batch size of 512 on 4 A40 GPUs. More details are provided in Appendix B.1.

Results. Consistent with fVLM [6], we report the macro area under ROC curve (AUC); balanced accuracy (ACC); weighted F1-score (F1); recall; and precision for multi-label zero-shot classification in Table 1. When pre-trained with original reports, HLIP outperforms second-best models by 4.9% AUC on internal and 7.9% AUC on external validation. When pre-trained with reports summarized by large language models (e.g., Qwen [52], as used by fVLM), HLIP outperforms fVLM by 0.9% and 3.7% AUC on internal and external validation, respectively. The external generalizability of fVLM is largely limited by its reliance on anatomical segmentation, which, while facilitating the learning of fine-grained representations, introduces quality-related bias. HLIP, by contrast, demonstrates that a simple adaptation of the ViT backbone is sufficient to achieve superior performance.

During implementation, we observe that reports summarized by Qwen can improve performance on internal validation (*i.e.*, +1.0% AUC), but do not necessarily lead to better performance on external validation (*i.e.*, -0.6% AUC). This observation suggests that overly summarized supervision signals may actually constrain generalizability. This further prompted us to investigate the question: *What additional factors could influence language-image pre-training in radiology*? As this analysis lies beyond the main focus of the paper, we present the results in Appendix C.

4.2 Brain MRI

Pub-Brain-5. To the best of our knowledge, no publicly available benchmark exists for zero-shot transfer evaluation on brain MRI tasks.[2] To this end, we construct a benchmark, named *Pub-Brain-5*, based on existing publicly available brain MRI datasets. Pub-Brain-5 comprises 18,343 studies drawn from Open-BHB [45], the Stroke dataset [21], BraTS 2023 [46, 47, 48, 49], NYU-Mets [51], and UCSF-Mets [50]. It spans five classes: healthy (3,984), acute stroke (2,871), glioma (1,614), meningioma (1,141), and metastasis (8,733). For each study, the number of scans ranges from 1 to 14. We further construct a subset of Pub-Brain-5, namely *Pub-Brain-5-GT*, which includes segmentation ground truth from the original datasets, enabling compatibility with medical foundation models that require 2D inputs [23, 24]. Pub-Brain-5-GT comprising 8,944 studies drawn from Open-BHB, the Stroke dataset, and BraTS 2023, covering the same five classes: healthy (3,984), acute stroke (2,372), glioma (1,350), meningioma (1,000), and metastasis (238). For both benchmarks, we evaluate three tasks: (i) binary disease detection (*i.e.*, distinguishing diseased from healthy studies); (ii) three-class brain tumor classification; and (iii) five-class brain disease classification.

---

[2]BrainMD [37] is not publicly available due to an unexpected privacy policy.

Table 2: Results of zero-shot classification on Pub-Brain-5-GT and Pub-Brain-5. We report balanced accuracy (multi-class), with the best results highlighted in **bold**. "+*annotation*" indicates only using lesion-containing slices which are manually annotated by expert radiologists. These values are an upper-bound scenario for baselines, but *not practical in real-world clinical settings*.

Method	Disease Detection				Tumor Classification	Disease Classification
	Stroke	Glioma	Meningioma	Metastasis		
Pub-Brain-5-GT						
BiomedCLIP [23]	66.7	88.2	63.8	74.6	46.2	33.1
+*annotation*	86.4	94.1	75.8	75.0	45.7	45.3
ConceptCLIP [24]	69.6	92.1	57.8	69.5	35.2	31.6
+*annotation*	93.6	**97.8**	70.8	**76.8**	39.4	50.8
HLIP (ours)	**95.0**	89.2	**79.6**	73.4	**54.8**	**61.3**
Pub-Brain-5						
BiomedCLIP [23]	64.7	87.8	63.6	59.8	50.4	31.5
ConceptCLIP [24]	66.8	**91.9**	57.7	67.9	35.7	30.9
HLIP (ours)	**91.5**	89.2	**79.2**	**78.1**	**63.3**	**63.9**

Baselines. Given that recent 3D models such as Merlin [4] and M3D [5] do not include brain MRI in their training sets, we evaluate two 2D foundation models on our benchmark: BiomedCLIP [23], pre-trained on *15 million* figure-caption pairs from PubMed, and ConceptCLIP [24], pre-trained on *23 million* such pairs. We use the prompt "*This brain MRI shows: {disease}.*" to perform zero-shot inference [2] with both models. Since both models require 2D inputs, we generate predictions for each study by applying *average pooling* and *max pooling* to the outputs across all slices. We observe that max pooling consistently yields positive predictions and is only suitable for binary classification; accordingly, we adopt average pooling for both models. Additionally, on Pub-Brain-5-GT, we assess an upper-bound scenario in which predictions are made only on lesion-containing slices. Note that this setting is not practical, as it requires manually annotation for the lesion locations.

Implementation Details. HLIP is trained on BrainMRI220K as described in Section 3.2, using the prompt "*This MRI study shows: {disease}.*" during zero-shot transfer [2].

Results. As two benchmarks are imbalanced, we report balanced accuracy for multi-class classification (mACC) in Table 2. On Pub-Brain-5-GT, we observe a performance boost for 2D foundation models when predicting on lesion-containing slices. Nevertheless, even when compared with the upper-bound scenario of baselines, HLIP still demonstrates superior performance in zero-shot brain tumor classification, *i.e.*, outperforming BiomedCLIP [23], by 8.6% mACC; and in brain disease classification, *i.e.*, outperforming ConcepCLIP [24], by 10.5% mACC. When evaluated on the more comprehensive Pub-Brain-5 dataset, the performance gap for 2D foundation models widens due to the absence of lesion-containing annotations. Specifically, HLIP outperforms the second-best models by 12.9% mACC in brain tumor classification, and by 32.4% mACC in disease classification.

HLIP can occasionally lag behind BiomedCLIP [23] and ConceptCLIP [24], particularly in the detection of brain glioma. However, we note that datasets like BraTS [46] are likely included in their pre-training corpora, as both models are trained on figure-caption pairs from PubMed. HLIP, on the other hand, does not encounter any instances from public datasets during training. The skull-stripping used in BraTS actually exacerbates the distribution shift. Moreover, HLIP makes predictions directly from the original radiology study, offering greater flexibility for real-world applications.

4.3 Head CT

Implementation Details. We apply HLIP to HeadCT240K, as described in Section 3.2. We then perform linear probing evaluation on two public benchmarks, RSNA [26] and CQ500 [27], exactly following FM-HeadCT [25], which is trained on 361,633 head CT scans with DINOv2 [53].

Results. Consistent with FM-HeadCT [25], we report the area under the ROC curve (AUC) in Figure 3. On the RSNA dataset [26], HLIP outperforms the second-best model, FM-HeadCT [25], on 4 out of 5 tasks, achieving an average AUC improvement of 1.4%. On the CQ500 dataset [27], HLIP outperforms both FM-HeadCT and Google-CT [54] on 7 out of 10 tasks, with average AUC improvements of 6.9% and 10.8%, respectively.

(a) RSNA head CT hemorrhage dataset benchmarks a 5-way intracranial hemorrhage diagnosis task.

(b) CQ500 dataset benchmarks a 10-way task that includes hemorrhage diagnosis, bleed localization, midline shift, and mass effect detection.
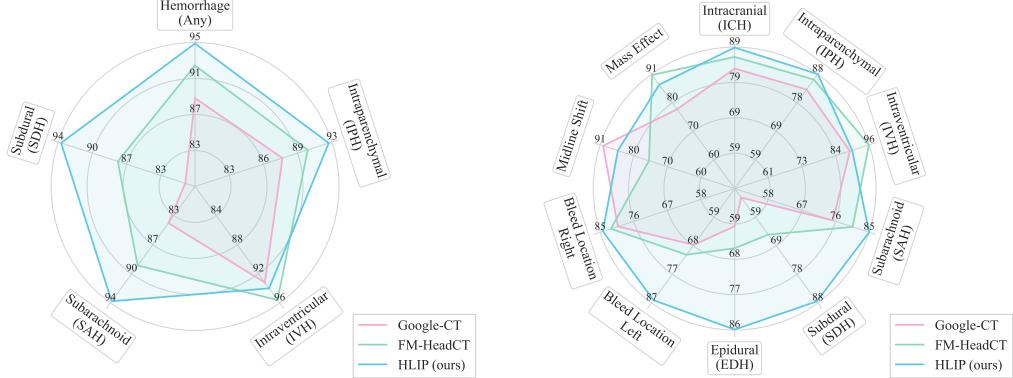
Figure 3: Results of linear probing on the (a) RSNA [26] and (b) CQ500 [27] datasets. We report AUC, with our method shown in blue, FM-HeadCT [25] in green, and Google CT [54] in red.
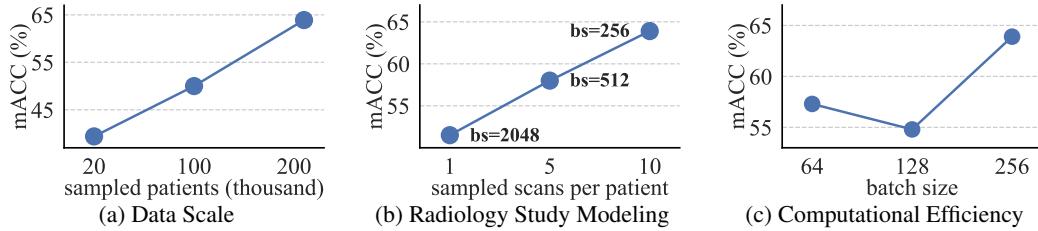

(a) Data Scale

(b) Radiology Study Modeling

(c) Computational Efficiency

Figure 4: Results of disease classification on Pub-Brain-5, highlighting the equal importance of (a) data scale, (b) radiology study modeling, and (c) computational efficiency. We report mACC.

(a) cls token strategy

(b) scan attn location

(c) # scan attn

(d) batch size

(e) cls token strategy
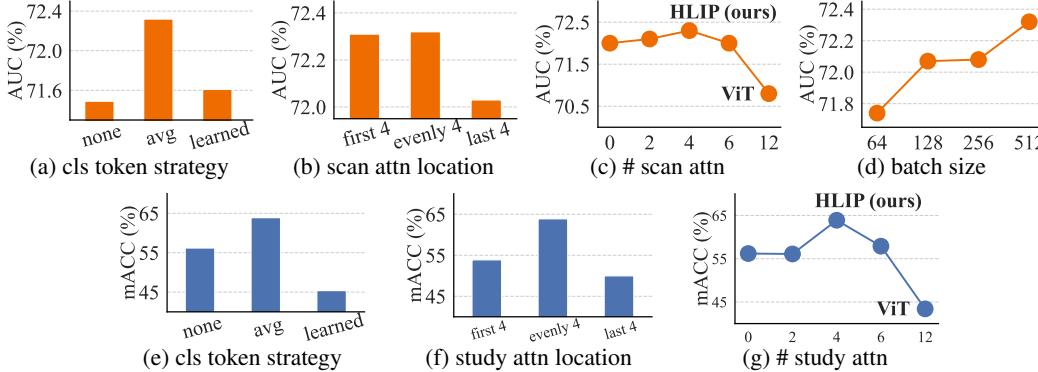
(f) study attn location

(g) # study attn

Figure 5: Ablation study on the Rad-ChestCT [22] dataset (a)-(d) and the Pub-Brain-5 dataset (e)-(g). We report AUC and mACC for these two benchmarks, respectively.

4.4 Scaling, Modeling, and Efficiency

We acknowledge that one reason HLIP achieves superior performance (*e.g.*, on Pub-Brain-5) is its training on a large-scale, domain-specific dataset. As shown in Figure 4a, zero-shot disease classification on Pub-Brain-5 drops by 24.5% mACC when using only 10% of the training data. However, data scale is not the only contributing factor. Given an uncurated dataset, a straightforward approach is to randomly select one scan per study *per step* while training a 3D ViT. Modeling radiology studies in this way results in a 12.4% drop in mACC, even when trained on the full dataset with the same computational resources, as shown in Figure 4b. Computational efficiency is equally important, as it determines how many samples can be contrasted within a batch. As shown in Figure 4c, a small batch size (e.g., *64*) results in a 6.6% drop in mACC. Therefore, HLIP bridges large-scale data, effective study modeling, and computational efficiency, enabling it to outperform current state-of-the-art foundation models [23, 24] by a substantial margin.
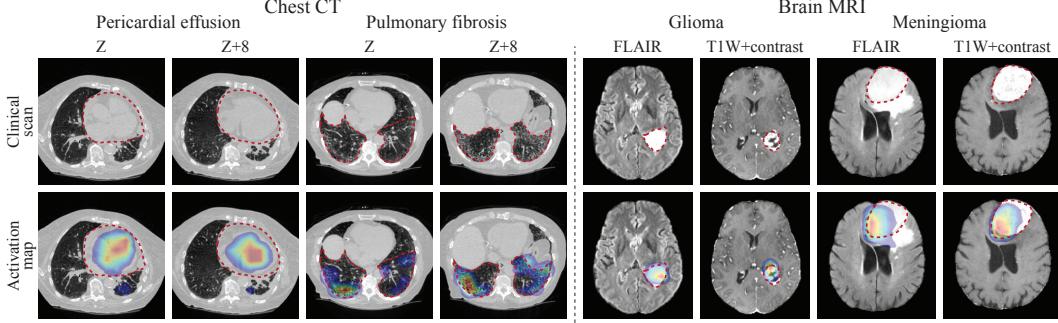
8

Figure 6: Qualitative results of zero-shot diagnosis on the Rad-ChestCT [22] and BraTS [46, 47] datasets. The first row shows the original clinical scans with pathologic regions outlined (dashed red). The second row shows activation HLIP maps [55]. HLIP identifies the pathologic regions across multiple groups of adjacent chest CT slides (left) and brain MRI scans (right).

4.5 Ablations

Ablation on Rad-ChestCT. For the Rad-ChestCT dataset, our visual encoder is composed of slice attention layers with 4 evenly distributed scan attention layers. In Figure 5a, we first ablate our strategy for propagating the *cls token* throughout the encoder, where *none* refers to applying average pooling at the end of the encoder, and *learned* refers to aggregate multiple *cls tokens* using attention pooling. We ablate the location of the scan attention layers in Figure 5b, and the number of scan attention layers in Figure 5c, where using 12 scan layers corresponds to the original ViT model—further highlighting the effectiveness of our hierarchical attention mechanism compared to ViT. In addition, we show that large batch size can achieve better performance in Figure 5d.

Ablation on Pub-Brain-5. For the Pub-Brain-5 dataset, the encoder is composed of scan attention layers with 4 evenly distributed study attention layers. We ablate the *cls token*, the location and the number of study attention layers in Figure 5e, Figure 5f, and Figure 5g, respectively. Figure 5g again highlights the effectiveness of our modeling compared to the original ViT.

4.6 Visualizations and Clinical Translation

Visualizations. In Figure 6, we visualize the activation maps [55] of HLIP in the zero-shot setting on Rad-ChestCT [22] and BraTS [46, 47]. For chest CT (curated, single scan), HLIP attends to pathologic regions across different slices, e.g., slice Z and Z+8, by leveraging scan attention. For brain MRI (uncurated, multiple scans), HLIP attends to pathologic regions across different scan types, e.g., FLAIR and T1W+contrast, by leveraging study attention. This demonstrates the effectiveness HLIP in modeling the hierarchical structure of 3D medical imaging.

Clinical Translation of HLIP. We conduct a 1-year, health system-scale, comprehensive evaluation on a prospective set of neurological patients. The datasets include ~23K brain MRI studies covering 52 diagnoses and ~15K head CT studies covering 83 diagnoses. We report macro AUC in Figure 7a and Figure 7b. HLIP consistently outperforms ViT. We provide detailed results in Appendix D.

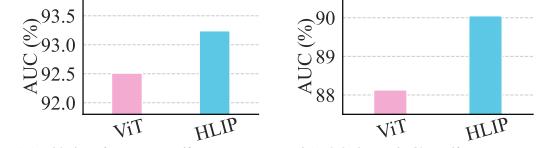
(a) 52 brain MRI diagnoses   (b) 83 head CT diagnoses

Figure 7: Prospective testing results on (a) 52 brain MRI diagnoses and (b) 83 head CT diagnoses. Task-wise results are presented in Figures 12 and 13 in Appendix.

5 Conclusion

In this work, we introduce HLIP, a scalable language-image pre-training framework for 3D medical imaging. By leveraging the inherent hierarchical structure of radiology studies, HLIP enables pre-training directly on uncurated CT and MRI datasets at health system scale. Across diverse modalities and anatomical regions, including chest CT, head CT, and brain MRI, HLIP achieves state-of-the-art performance on multiple benchmarks. These results demonstrate that, with HLIP, directly pre-training on uncurated clinical datasets is a scalable and effective direction for language-image pre-training in 3D medical imaging. We hope our work facilitates the language-image pre-training in other health systems and inspires future research on scalable approaches for real-world clinical data.

References

[1] Yuhao Zhang et al. "Contrastive learning of medical visual representations from paired images and text". In: *Machine learning for healthcare conference*. PMLR. 2022, pp. 2–25.

[2] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

[3] Ibrahim Ethem Hamamci et al. "Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography". In: *arXiv preprint arXiv:2403.17834* (2024).

[4] Louis Blankemeier et al. "Merlin: A Vision Language Foundation Model for 3D Computed Tomography". In: *arXiv preprint arXiv:2406.06512* (2024).

[5] Fan Bai et al. "M3d: Advancing 3d medical image analysis with multi-modal large language models". In: *arXiv preprint arXiv:2404.00578* (2024).

[6] Zhongyi Shui et al. "Large-scale and Fine-grained Vision-language Pre-training for Enhanced CT Image Understanding". In: *arXiv preprint arXiv:2501.14548* (2025).

[7] Benedikt Boecking et al. "Making the most of text semantics to improve biomedical vision–language processing". In: *European conference on computer vision*. Springer. 2022, pp. 1–21.

[8] Zifeng Wang et al. "Medclip: Contrastive learning from unpaired medical images and text". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*. Vol. 2022. 2022, p. 3876.

[9] Ekin Tiu et al. "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning". In: *Nature biomedical engineering* 6.12 (2022), pp. 1399–1406.

[10] Kihyun You et al. "Cxr-clip: Toward large scale chest x-ray language-image pre-training". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 101–111.

[11] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2020.

[12] Kaiming He et al. "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.

[13] Mehdi Cherti et al. "Reproducible scaling laws for contrastive language-image learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 2818–2829.

[14] Ze Liu et al. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[15] Haoqi Fan et al. "Multiscale vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 6824–6835.

[16] Yanghao Li et al. "Mvitv2: Improved multiscale vision transformers for classification and detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 4804–4814.

[17] Chaitanya Ryali et al. "Hiera: A hierarchical vision transformer without the bells-and-whistles". In: *International conference on machine learning*. PMLR. 2023, pp. 29441–29454.

[18] Yanghao Li et al. "Exploring plain vision transformer backbones for object detection". In: *European conference on computer vision*. Springer. 2022, pp. 280–296.

[19] Tri Dao et al. "Flashattention: Fast and memory-efficient exact attention with io-awareness". In: *Advances in neural information processing systems* 35 (2022), pp. 16344–16359.

[20] Yanghao Li et al. "Scaling language-image pre-training via masking". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 23390–23400.

[21] Chin-Fu Liu et al. "A large public dataset of annotated clinical MRIs and metadata of patients with acute stroke". In: *Scientific Data* 10.1 (2023), p. 548.

[22] Rachel Lea Draelos et al. "Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes". In: *Medical image analysis* 67 (2021), p. 101857.

[23] Sheng Zhang et al. "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs". In: *arXiv preprint arXiv:2303.00915* (2023).

[24]   Yuxiang Nie et al. "ConceptCLIP: Towards Trustworthy Medical AI via Concept-Enhanced Contrastive Langauge-Image Pre-training". In: *arXiv preprint arXiv:2501.15579* (2025).

[25]   Weicheng Zhu et al. "3D Foundation AI Model for Generalizable Disease Detection in Head Computed Tomography". In: *arXiv preprint arXiv:2502.02779* (2025).

[26]   Adam E Flanders et al. "Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge". In: *Radiology: Artificial Intelligence* 2.3 (2020), e190211.

[27]   Sasank Chilamkurthy et al. "Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study". In: *The Lancet* 392.10162 (2018), pp. 2388–2396.

[28]   Jeremy Irvin et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597.

[29]   Alistair EW Johnson et al. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". In: *Scientific data* 6.1 (2019), p. 317.

[30]   Shih-Cheng Huang et al. "Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 3942–3951.

[31]   Fuying Wang et al. "Multi-granularity cross-modal alignment for generalized medical visual representation learning". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 33536–33549.

[32]   Philip Müller et al. "Joint learning of localized representations from medical images and reports". In: *European conference on computer vision*. Springer. 2022, pp. 685–701.

[33]   Chaoyi Wu et al. "Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 21372–21383.

[34]   Shruthi Bannur et al. "Learning to exploit temporal structure for biomedical vision-language processing". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15016–15027.

[35]   Xiaoman Zhang et al. "Knowledge-enhanced visual-language pre-training on chest radiology images". In: *Nature Communications* 14.1 (2023), p. 4542.

[36]   Weiwei Cao et al. "Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 11238–11247.

[37]   Yuli Wang et al. "Enhancing vision-language models for medical imaging: bridging the 3D gap with innovative slice selection". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 99947–99964.

[38]   Xiaoxuan He et al. "Unified Medical Image Pre-training in Language-Guided Common Semantic Space". In: *European Conference on Computer Vision*. Springer. 2024, pp. 123–139.

[39]   Haoran Lai et al. "Bridged Semantic Alignment for Zero-shot 3D Medical Image Diagnosis". In: *arXiv preprint arXiv:2501.03565* (2025).

[40]   Christos Matsoukas et al. "What makes transfer learning work for medical images: Feature reuse & other factors". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 9225–9234.

[41]   Chenhui Zhao and Liyue Shen. "Part-aware Personalized Segment Anything Model for Patient-Specific Segmentation". In: *arXiv preprint arXiv:2403.05433* (2024).

[42]   Alok Aggarwal and S Vitter Jeffrey. "The input/output complexity of sorting and related problems". In: *Communications of the ACM* 31.9 (1988), pp. 1116–1127.

[43]   Yuhui Zhang et al. "Adapting pre-trained vision transformers from 2d to 3d through weight inflation improves medical image segmentation". In: *Machine Learning for Health*. PMLR. 2022, pp. 391–404.

[44]   Yu Gu et al. *Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing*. 2020. eprint: `arXiv:2007.15779`.

[45]   Benoit Dufumier et al. "Openbhb: a large-scale multi-site brain mri data-set for age prediction and debiasing". In: *NeuroImage* 263 (2022), p. 119637.

[46]   Ujjwal Baid et al. "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification". In: *arXiv preprint arXiv:2107.02314* (2021).

[47] Dominic LaBella et al. *The ASNR-MICCAI Brain Tumor Segmentation (BraTS) Challenge 2023: Intracranial Meningioma*. 2023. arXiv: 2305.07642 [cs.CV].

[48] Ahmed W Moawad et al. "The Brain Tumor Segmentation-Metastases (BraTS-METS) Challenge 2023: Brain Metastasis Segmentation on Pre-treatment MRI". In: *ArXiv* (2024), arXiv–2306.

[49] Anahita Fathi Kazerooni et al. "The brain tumor segmentation (BraTS) challenge 2023: focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs)". In: *ArXiv* (2024), arXiv–2305.

[50] Jeffrey D Rudie et al. "The University of California San Francisco Brain Metastases Stereotactic Radiosurgery (UCSF-BMSR) MRI Dataset". In: *Radiology: Artificial Intelligence* 6.2 (2024), e230126.

[51] Katherine E Link et al. "Longitudinal deep neural networks for assessing metastatic brain cancer on a large open benchmark". In: *Nature Communications* 15.1 (2024), p. 8170.

[52] Jinze Bai et al. "Qwen technical report". In: *arXiv preprint arXiv:2309.16609* (2023).

[53] Maxime Oquab et al. "Dinov2: Learning robust visual features without supervision". In: *arXiv preprint arXiv:2304.07193* (2023).

[54] Lin Yang et al. "Advancing multimodal medical capabilities of Gemini". In: *arXiv preprint arXiv:2405.03162* (2024).

[55] Hila Chefer, Shir Gur, and Lior Wolf. "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 397–406.

[56] Bjoern H Menze et al. "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024.

[57] Spyridon Bakas et al. "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features". In: *Scientific data* 4.1 (2017), pp. 1–13.

[58] Jun Ma et al. "Segment anything in medical images". In: *Nature Communications* 15.1 (2024), p. 654.

[59] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature methods* 18.2 (2021), pp. 203–211.

[60] Samuel Joutard, Maximilian Pietsch, and Raphael Prevost. "HyperSpace: Hypernetworks for spacing-adaptive image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 339–349.

[61] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

[62] Ilya Loshchilov and Frank Hutter. "Sgdr: Stochastic gradient descent with warm restarts". In: *arXiv preprint arXiv:1608.03983* (2016).

[63] M Jorge Cardoso et al. "Monai: An open-source framework for deep learning in healthcare". In: *arXiv preprint arXiv:2211.02701* (2022).

Appendix

A Dataset

As mentioned in Section 3.2, this section provides additional details on the two datasets we collected: BrainMRI220K and HeadCT240K. For other datasets, such as CT-RATE [3], Rad-ChestCT [22], BraTS 2023 [46, 47, 48, 49, 56, 57], NYU-Mets [51], UCSF-Mets [50], RSNA [26], and CQ500 [27], we refer readers to their original publications. Additionally, we include a discussion of our preprocessing strategy.

A.1 BrainMRI220K

(a) # scans per study   (b) # slices per scan   (c) # scans per acqusition plane
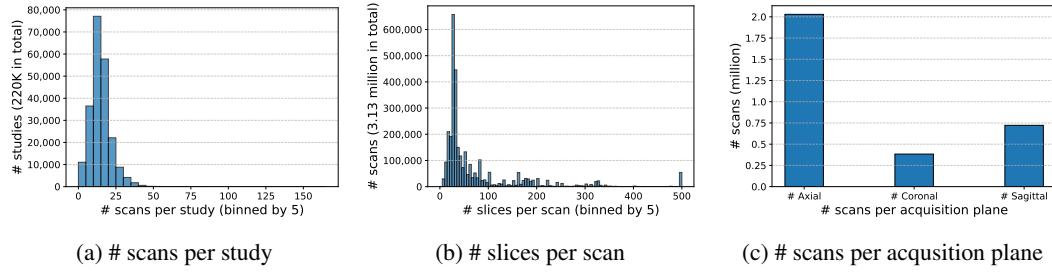
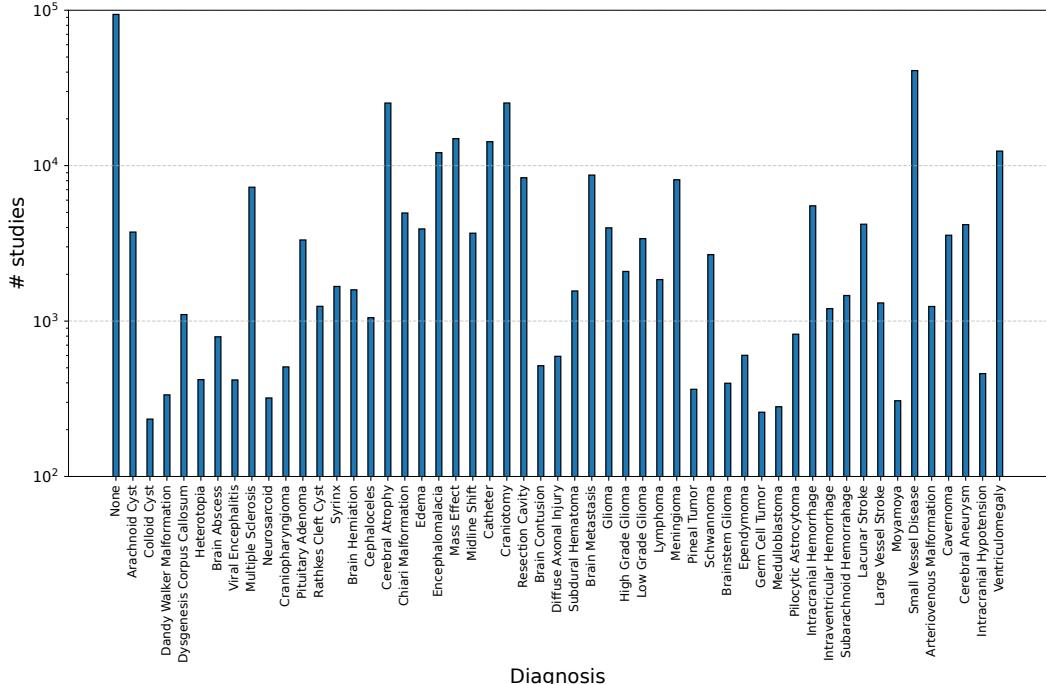Figure 8: Statistic results on BrainMRI220K.

Figure 9: Diagnosis distribution on BrainMRI220K.

13

The distributions of the number of scans per study, slices per scan, and scans per acquisition view in the BrainMRI220K dataset are shown in Figure 8. Specifically, the number of scans per study ranges from 1 to 162, with the third quartile at 17. In total, there are 3.13 million scans from 220,001 MRI studies. All scans are acquired with through-plane spacing equal to or smaller than 4mm. The number of slices per scan ranges from 5 to 500, with the third quartile at 80. Each scan is originally acquired from a different view, with the ratio between axial, coronal, and sagittal views being $5:1:2$. Moreover, we show the distribution of 52 brain MRI diagnoses in Figure 9, based on keyword statistics extracted from the reports.

A.2 HeadCT240K

(a) # scans per study   (b) # slices per scan   (c) # scans per acqusition plane
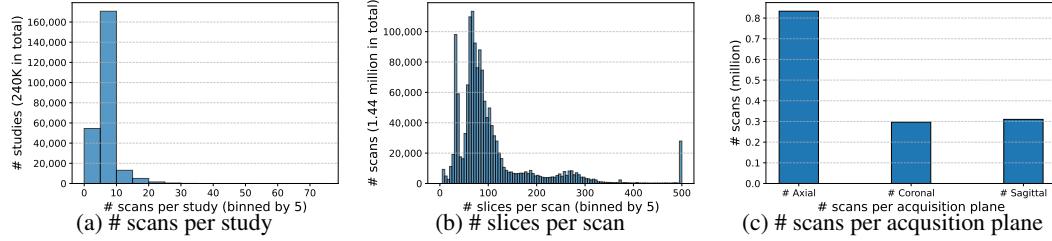
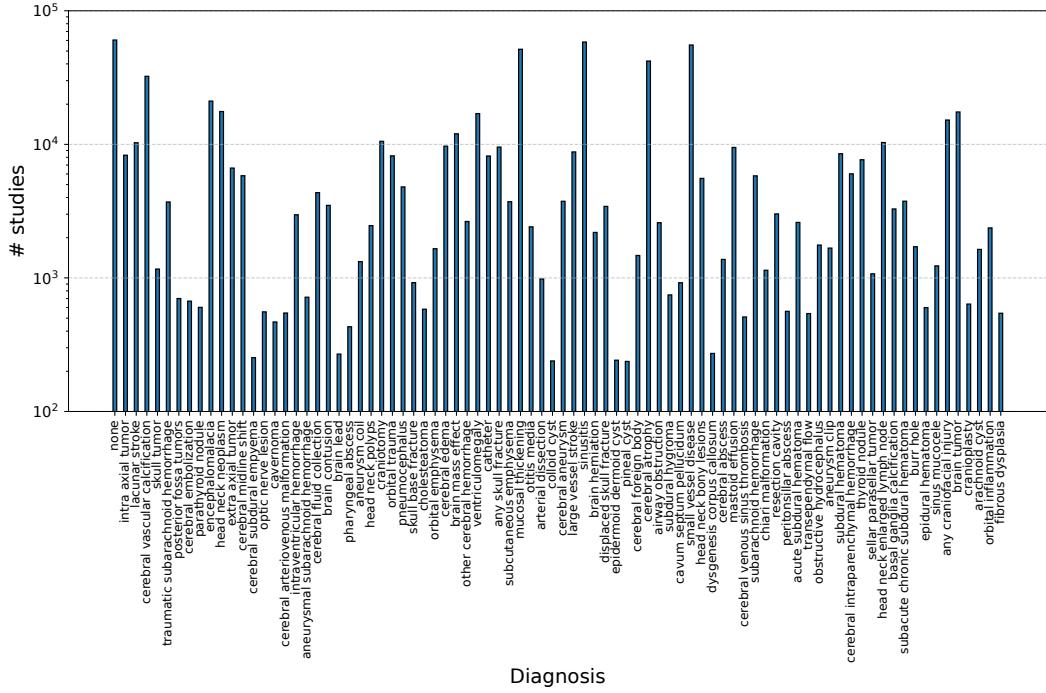Figure 10: Statistic results on HeadCT240K.

Figure 11: Diagnosis distribution on HeadCT240K.

The distributions of the number of scans per study, slices per scan, and scans per acquisition view in the HeadCT240K dataset are shown in Figure 10. Specifically, the number of scans per study ranges from 1 to 71, with the third quartile at 6. In total, there are 1.44 million scans from 244,253 CT studies. All scans are acquired with through-plane spacing equal to or smaller than 4mm. The number of slices per scan ranges from 5 to 500, with the third quartile at 110. Each scan is originally acquired from a different view, with the ratio between axial, coronal, and sagittal views being $8:3:3$. Moreover, we show the distribution of 83 head CT diagnoses in Figure 11, based on keyword statistics extracted from the reports.

14

A.3 Discussion

To the best of our knowledge, this is the first work to handle 3D medical datasets at the scale of BrainMRI220K and HeadCT240K, which comprise millions of scans. Our preprocessing strategy, introduced in Section 3.2, follows standard practices where possible, but also diverges due to the scale and diversity of our dataset. Specifically, percentile clipping for brain MRI and HU value truncation for head CT follow standard practices [25, 58]. For head CT, we do not concatenate scans with different HU truncations along the channel dimension, as our method is capable of processing multiple scans simultaneously. However, our strategy regarding orientation and spacing may differ from established preprocessing pipelines [3, 4, 6, 59]. Here, we provide a discussion to justify the rationale behind our approach.

Orientation. CT and MRI scans are acquired using different planes, including axial, coronal, and sagittal. Isotropic data are rare in routine radiology studies; the through-plane spacing is typically larger than the in-plane spacing. In the context of batch construction, where all samples in a batch must share the same shape, standardizing the orientation can lead to downsampling along the in-plane axes and upsampling along the through-plane axis. Upsampling in pixel space does not introduce new information, whereas downsampling results in information loss. From the perspective of data augmentation, the orientation can change when 3D random rotation is applied [59]. So, *why do we standardize the orientation at the beginning then augment it during training*? Instead, we consistently align the first dimension (depth) with the through-plane axis. Please note that the in-plane orientation is still standardized as RP, RI, or PI. We treat the diversity of acquisition planes (as shown in Figure 8c and Figure 10c) in our dataset as a form of natural data augmentation.

Spacing. Standardizing spacing across the entire dataset is a common practice in medical image segmentation to ensure that convolution filters interpret anatomical structures consistently [59]. However, HyperSpace [60] demonstrates that, with appropriate spacing augmentation, convolutional networks can learn spacing variations. Therefore, architectures like Vision Transformers, which can learn scale-equivariant features [11], should also be capable of learning spacing-equivariant features with sufficient spacing augmentation. This is because spacing equivariance is a direct extension of scale equivariance from 2D to 3D space. Given the diversity of spacing settings (as shown in Figure 8b and Figure 10b) in our dataset, we believe it naturally occupies sufficient spacing augmentation for the model to learn spacing-equivariant features. Therefore, when constructing a batch, we first resize each volume to $(48,256,256)$, and then apply a center crop to obtain $(48,224,224)$.

M **scans per study.** As the number of scans per study varies, we randomly sample $M = 10$ scans per study at each training step to construct a batch. An equal number of positional embeddings is also sampled for these M scans from a total of $M_{\max} = 40$ positional embeddings at each step. This means the model is exposed to $M_{\max}$ different positions over the course of training, while observing M positions at each step. This is analogous to a dropout strategy, with a scan dropout rate of $1 - \frac{M}{M_{\max}}$. During evaluation, the model is able to process studies containing up to $M_{\max}$ scans without the need for interpolation. We investigate the effects of M and $M_{\max}$ in Figure 4b, but we believe these hyperparameters should be tuned according to the distribution of the dataset.

B Implementation

In this section, we summarize the model configuration described in Section 3.2 and provide details of the pre-training setup referenced in both Section 3.2 and Section 4.1. For linear probing evaluation on head CT, we refer readers to FM-HeadCT [25].

B.1 Chest CT

The model architecture and pre-training configuration for chest CT are summarized in Table 3 and Table 4, respectively. These configurations closely follows prior works such as CT-CLIP [3], OpenCLIP [13], and BiomedCLIP [23]. In the pre-training configuration, we report the base learning rate corresponding to a batch size of 64. During training, we follow the linear learning rate scaling rule [20]: $lr = base\_lr \times \frac{batch\_size}{64}$.

The diseases used for zero-shot evaluation are listed in Table 5. This setup is consistent with fVLM [6] and all methods reported in Table 1.

Table 3: Model (Chest CT).

config	value
projection	linear
embed dim	512
Visual Encoder	ViT-B [11]
slice attn index	(0,1;3,4;6,7;9,10)
scan attn index	(2,5,8,11)
input size	(112,336,336)
patch size	(8,24,24)
patch dropout	0.0
Text Encoder	CXR-BERT [7]
context length	512 [3]

Table 4: Pre-training (Chest CT).

config	value
image precision	float32
truncate (window)	[-1150,350]
normalize (mean, std)	[0.449,0.226]
report	original
optimizer	AdamW [61]
base learning rate	1e-5
weight decay	0.2
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.98$ [2]
learning rate schedule	cosine decay [62]
warmup (in steps)	47 [6]
numerical precision	amp

Table 5: Evaluation (Chest CT).

CT-RATE [3]	Rad-ChestCT[22]
Emphysema	emphysema
Atelectasis	atelectasis
Lung nodule	nodule
Lung opacity	opacity
Pulmonary fibrotic sequela	fibrosis
Pleural effusion	pleural_effusion
Mosaic attenuation pattern	∅
Peribronchial thickening	bronchial_wall_thickening
Consolidation	consolidation
Bronchiectasis	bronchiectasis
Interlobular septal thickening	septal_thickening
Cardiomegaly	cardiomegaly
Pericardial effusion	pericardial_effusion
Coronary artery wall calcification	calcification
Hiatal hernia	hernia
Arterial wall calcification	calcification

B.2 Brain MRI/Head CT

Table 6: Model (Brain MRI/Head CT).

config	value
projection	linear
embed dim	512
Visual Encoder	ViT-B [11]
scan attn index	(0,1;3,4;6,7;9,10)
study attn index	(2,5,8,11)
input size	(48,224,224)
patch size	(8,16,16)
patch dropout	0.25
Text Encoder	PubMedBERT [23, 44]
context length	256 [23]

Table 7: Pre-training (Brain MRI/Head CT).

config	value
image precision	uint8
truncate (percentile)	[0.05,0.95]
normalize (mean, std)	[0.449,0.226]
report	GPT-3.5
optimizer	AdamW [61]
base learning rate	1e-4
weight decay	0.2
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$ [20]
learning rate schedule	cosine decay [62]
warmup (in steps)	2000
numerical precision	amp

The model architecture and pre-training configuration for brain MRI/head CT are summarized in Table 6 and Table 7, respectively. These configurations also follows prior works [13, 20, 23]. In the pre-training configuration, we report the base learning rate corresponding to a batch size of 128. During training, we follow the linear learning rate scaling rule [20]: $lr = base\_lr \times \frac{batch\_size}{128}$.

Table 8: Ablation studies of training differences between CT-CLIP [3] and HLIP on the CT-RATE [3] test split for internal validation and the Rad-ChestCT [22] for external validation.

Method	Internal Validation (CT-RATE)			External Validation (Rad-ChestCT)		
	AUC	ACC	F1	AUC	ACC	F1
CT-CLIP [3]	73.3	66.9	70.8	63.3	59.9	64.7
ViT (*our impl; bs=64*)	+2.0	+2.8	+2.4	–0.8	–0.8	–0.6
+ *IN.MAE*	+0.8	+0.1	+0.2	+7.3	+6.3	+5.4
+ *lock text*	+0.4	+0.5	+0.3	+1.2	+0.3	+0.2
+ *hierarchical attn*	+0.3	+1.0	+0.8	+0.7	+1.1	+1.0
+*batch size 512*	+0.9	+0.1	+0.2	+0.6	+0.6	+1.4
= HLIP	77.7	71.4	74.7	72.3	68.4	72.1
report→*Qwen*	78.7	72.4	75.5	71.7	67.7	71.4
image precision→*uint8*	78.0	71.8	75.0	72.4	67.2	71.1
construct batch→*resize*	77.9	71.3	74.6	71.9	67.6	71.4
truncate→`[-1000,1000]`	77.6	71.1	74.4	71.4	66.9	70.8

C   Additional Ablation Studies on Chest CT

In this section, we address the question: *What factors could influence language-image pre-training in radiology?* under the constraint of a fixed data scale. Specifically, we evaluate the impact of several components added to the CT-CLIP baseline [3] on chest CT benchmarks [3, 22]. As shown in Table 8:

- **ViT.** Our implementation of the original ViT [11] overfits well on the CT-RATE [3] training set. Compared with CT-CLIP [3], our implementation achieves a +2% AUC gain in internal evaluation and a –0.8% AUC drop in external evaluation. This can be attributed to the effective implementation adopted from OpenCLIP [13].

- **+MAE**. MAE [12] pre-trained weight improves performance in external evaluation by a substantial margin of +7.3% AUC. This result highlights the universal features learned by MAE, as well as the advantage of using the original ViT backbone.

- **+Lock Text Encoder.** During implementation, we observed severe overfitting even in the internal evaluation. Although the visual encoder is pre-trained with MAE, there exists a significant domain gap. In contrast, the text encoder is CXR-BERT [7], whose training domain, chest X-ray diagnosis, substantially overlaps with chest CT. We hypothesize that the overfitting is primarily caused by the text encoder and thus freeze it during training. This strategy alleviates the overfitting issue and yields a +1.2% improvement in AUC.

- **HLIP (bs=512).** Our HLIP further improves performance in both internal and external evaluations, achieving over 1% improvement across all three metrics, benefiting from both effective modeling and computational efficiency.

It is important to note that the performance improvement brought by HLIP being smaller than that of the MAE pre-trained weights does not imply that HLIP's advantage lies solely in using MAE initialization. fVLM [6] also incorporates MAE pre-trained weights but still lags behind HLIP in both internal and external evaluations, as shown in Section 4.1. Therefore, it is the effectiveness of the complete HLIP framework that accounts for the observed improvement.

We additionally evaluate other factors that may be of interest to researchers during implementation. Among these, the most informative is saving the data in the *uint8* format. Since language-image pre-training in 3D medical imaging requires loading volumetric data, data loading I/O can become a training-time bottleneck. As storing data in *uint8* format does not affect performance, it offers a more efficient solution for both data loading and storage. Other factors, such as supervision summarized by Qwen [52], using resize operation instead of standardizing the spacing when constructing the batch, or different HU value truncation, only lead to subtle variations in internal evaluation. In external evaluation, we believe the performance variation is primarily caused by overfitting, which is expected to diminish as training scales up.

Therefore, identifying an effective strategy to scale up language-image pre-training in 3D medical imaging may be the most critical factor. Other design choices should be adopted insofar as they facilitate data scalability.
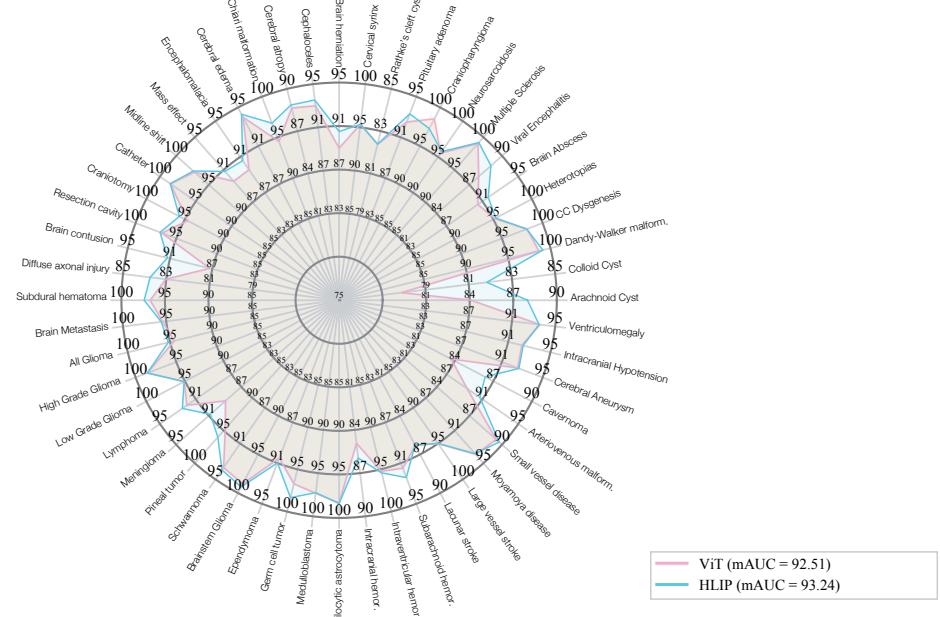
D Prospective Evaluation

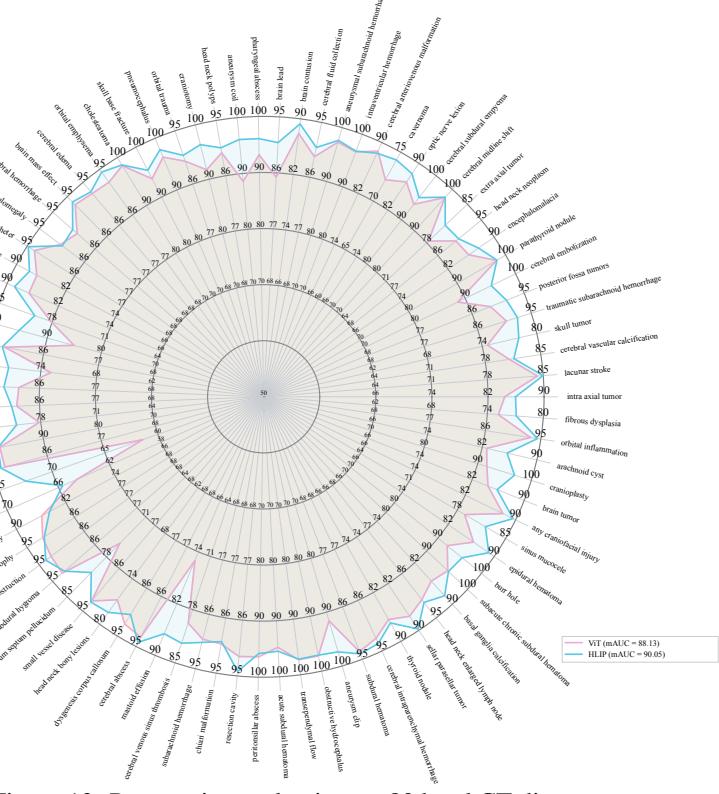Figure 12: Prospective evaluation on 52 brain MRI diagnoses.

Figure 13: Prospective evaluation on 83 head CT diagnoses.

We present the detailed results of the prospective evaluation conducted within our health system for 52 brain MRI diagnoses and 83 head CT diagnoses in Figure 12 and Figure 13, respectively. We report linear probing AUC for HLIP (blue) and the original ViT model (red). HLIP demonstrate consistent improvement.

18

E Complexity

E.1 Analysis

Given a sequence length of N, model dimension of c, the matrices $Q, K, V \in \mathbb{R}^{N \times c}$. The output $O \in \mathbb{R}^{N \times c}$ of standard self-attention is computed as:

$$S = Q \times K^T \in \mathbb{R}^{N \times N}; \quad P = softmax(S) \in \mathbb{R}^{N \times N}; \quad O = PV \tag{1}$$

where S and P require $\Omega(N^2)$ memory access; Q, K, V, O requires $\Omega(N \times c)$ memory access. Therefore, the I/O complexity for a standard self-attention over N tokens is:

$$\Omega(N^2 + N \times c) \tag{2}$$

The S requires $\Omega(N^2 \times c)$ multiply-add operations; P requires $\Omega(N^2)$ multiply-add operations; O requires $\Omega(N^2 \times c)$ multiply-add operations. Therefore the compute complexity for a standard self-attention over N tokens is:

$$\Omega(N^2 \times c) \tag{3}$$

Study Attention. In our scenario, the sequence length of a study is $N = M \times d \times h \times w$. The study attention is a standard self-attention over all N tokens. Therefore the I/O complexity and the compute complexity of the study attention is $\Omega(N^2 + N \times c)$ and $\Omega(N^2 \times c)$, respectively, as shown in Equation 2 and Equation 3.

Scan Attention. The scan attention is M standard self-attention operations, each over $\frac{N}{M}$ tokens. Therefore replace N with $\frac{N}{M}$ in Equation 2 and Equation 3 resulting in $\Omega(\frac{N^2}{M^2} + \frac{N \times c}{M})$ and $\Omega(\frac{N^2}{M^2} \times c)$ for I/O and compute complexity of each self-attention operations, therefore the total I/O and compute complexity of the scan attention are $\Omega(\frac{N^2}{M} + N \times c)$ and $\Omega(\frac{N^2}{M} \times c)$, respectively.

Slice Attention. The slice attention is $M \times d$ standard self-attention operations, each over $\frac{N}{M \times d}$ tokens. Therefore, similar to the scan attention the total I/O and compute complexity of the slice attention are $\Omega(\frac{N^2}{M \times d} + N \times c)$ and $\Omega(\frac{N^2}{M \times d} \times c)$, respectively.

E.2 Experiments

We conduct experiments comparing ViT [11], the 3D Swin Transformer implemented in MONAI [63], and our HLIP visual encoder. Since Swin is not compatible with multi-scan inputs, the experiment is conducted on a single-channel 3D input of shape (B,1,224,224,224), where B=1 denotes the batch size. Both ViT and HLIP use a ViT-Base architecture with patch_size=(16,16,16). The Swin Transformer is configured with dim=128, patch_size=(4,4,4), window_size=(7,7,7), depth=(2,2,12,2), and head=(4,8,16,32). Following the original Swin Transformer [14] while matching the model size.

Table 9: Practical Complexity.

		w/o flash attn			w/ flash attn [19]	
	model size (M)	throughput (img/s) ↑=better	memory (G) ↓=better	model size (M)	throughput (img/s) ↑=better	memory (G) ↓=better
ViT	88.2	9.0	6.6	88.2	44.4	1.5
Swin	87.9	9.9	8.0	87.9	9.9	9.0
HLIP (ours)	88.2	17.5	4.1	88.2	47.6	1.5

All experiments are conducted on a single A40 GPU. We report the model size, throughput and training memory for each model. To measure throughput, we include 100 warm-up iterations and compute the average over 1000 forward passes. As shown in Table 9, when not using flash attention [19], Swin is slightly faster than the original ViT but still lags behind our HLIP visual encoder. Swin also consumes more memory during training due to the large codebook of relative position embeddings constructed for 3D inputs. Moreover, Swin is not compatible with flash attention, which has been integrated into PyTorch 2.0 and supports a wide range of GPUs. We acknowledge that when using flash attention, the memory efficiency advantage of our hierarchical attention is reduced, since we do not alter the actual feature map size. However, HLIP remains faster than the original ViT.

Note that the results shown in Table 9 differ from those in Figure 1, which are measured on uncurated studies with input size (B,8,1,48,224,224) and without flash attention.

F   Discussion

We propose HLIP, a scalable language-image pre-training framework for 3D medical imaging. In this section, we provide additional discussion of our approach.

Efficient Implementation. While we do not explicitly emphasize implementation efficiency in the paper, it is reflected in the training times reported in Section 3.2 and Section 4.1. To the best of our knowledge, training for 20 epochs on CT-RATE [3] within 6 hours using 4 A40 GPUs represents the most efficient implementations in this domain. To support further development in language-image pre-training for 3D medical imaging, we will release our implementation soon.

Limitations. Here, we further discuss three limitations of our work:

- The first limitation is computational resources. The most intensive compute setup used in this work is a node with 8 L40 GPUs, which is comparable to configurations used in other institutions. Although HLIP demonstrates substantial efficiency compared to the original ViT, we rely on both flash attention and gradient checkpointing to further optimize training. Nevertheless, the batch sizes achievable with our current resources are 256 for uncurated datasets and 512 for curated datasets which are still significantly smaller than the typical batch sizes used in the natural image and 2D medical imaging domains. Furthermore, our compute resources are not sufficient to train larger models such as ViT-Large.

- We also observe that zero-shot transfer performance does not always correlate with the number of studies. For example, although keyword search yields more patients with meningioma or metastasis than glioma, the zero-shot performance for glioma is substantially higher than for the other two. While tumor size may partially explain this discrepancy, we believe it warrants further investigation in future work.

- In this work, we simply collect all studies from our health system, resulting in an imbalanced pre-training dataset. Building on findings from the natural image domain, we believe that developing a systematic approach to construct a relatively balanced pre-training dataset is an important direction for future work.

Broader Impacts.   HLIP models pre-trained on our collected datasets, BrainMRI220K and HeadCT240K, will be released shortly. Since HLIP can make predictions directly on raw radiology studies, it offers flexibility for real-world clinical use and may serve as a reference tool for radiologists, potentially alleviating their workload. However, despite HLIP's strong performance on several benchmarks, it should not be solely relied upon for diagnostic decisions.