

---

# HSENet: Hybrid Spatial Encoding Network for 3D Medical Vision-Language Understanding

---

Yanzhao Shi<sup>1</sup>, Xiaodan Zhang<sup>2</sup>, Junzhong Ji<sup>2</sup>, Haoning Jiang<sup>1</sup>,  
Chengxin Zheng<sup>2</sup>, Yinong Wang<sup>1</sup>, Liangqiong Qu<sup>1</sup>

<sup>1</sup> The University of Hong Kong, Hong Kong

<sup>2</sup> Beijing University of Technology, Beijing, China

[zhangxiaodan@bjut.edu.cn](mailto:zhangxiaodan@bjut.edu.cn), [liangqqu@hku.hk](mailto:liangqqu@hku.hk)

## Abstract

Automated 3D CT diagnosis empowers clinicians to make timely, evidence-based decisions by enhancing diagnostic accuracy and workflow efficiency. While multi-modal large language models (MLLMs) exhibit promising performance in visual-language understanding, existing methods mainly focus on 2D medical images, which fundamentally limits their ability to capture complex 3D anatomical structures. This limitation often leads to misinterpretation of subtle pathologies and causes diagnostic hallucinations. In this paper, we present Hybrid Spatial Encoding Network (HSENet), a framework that exploits enriched 3D medical visual cues by effective visual perception and projection for accurate and robust vision-language understanding. Specifically, HSENet employs dual-3D vision encoders to perceive both global volumetric contexts and fine-grained anatomical details, which are pre-trained by dual-stage alignment with diagnostic reports. Furthermore, we propose Spatial Packer, an efficient multimodal projector that condenses high-resolution 3D spatial regions into a compact set of informative visual tokens via centroid-based compression. By assigning spatial packers with dual-3D vision encoders, HSENet can seamlessly perceive and transfer hybrid visual representations to LLM’s semantic space, facilitating accurate diagnostic text generation. Experimental results demonstrate that our method achieves state-of-the-art performance in 3D language-visual retrieval (39.85% of R@100, +5.96% gain), 3D medical report generation (24.01% of BLEU-4, +8.01% gain), and 3D visual question answering (73.60% of Major Class Accuracy, +1.99% gain), confirming its effectiveness. Our code is available at <https://github.com/YanzhaoShi/HSENet>.

## 1 Introduction

3D computed tomography (CT) has revolutionized medical diagnostics by providing high-resolution visualization of anatomical structures. Nonetheless, interpreting 3D CT images is labor-intensive for radiologists, which relies heavily on intricate psychophysiological and cognitive processes that are prone to perceptual errors [5]. The application of computer-aided diagnostic models offers considerable promise in assisting radiologists for efficient and accurate clinical decision-making.

Recently, multi-modal large language models (MLLMs) have emerged as a powerful tool in medical image analysis, including diagnostic tasks such as medical report generation (MRG) and visual question answering (VQA). Current works mainly focus on 2D medical imaging, such as X-ray [18, 41, 26, 37], which offers planar projections valuable for screening thoracic conditions and skeletal disorders. However, 2D imaging inherently fails to capture volumetric details of complex anatomical relationships, restricting the ability of MLLMs to interpret spatial patterns in lesions. This restriction hinders their clinical utility of models in scenarios requiring volumetric analysis, such as tumor

infiltration assessment or vascular anomaly detection. To address this challenge, early studies shift toward 3D CT imaging, employing slice-by-slice analysis [23, 49] or in chunks of small stacks of 2D slices [16], yet these methods still struggle to capture spatial continuity along the depth (z-axis) dimension. In contrast, RadFM [43] and M3D [3] leverage 3D Vision Transformers (ViTs) to train foundation MLLMs, utilizing a large volume of 3D medical samples to enhance the model adaptability across various tasks. To further reduce diagnostic hallucinations and improve clinical performance, these foundation models are integrated with specialized visual pretraining strategies [45, 21, 31] and visual encoding pipelines [36, 11, 6]. Nevertheless, existing methods still encounter challenges in understanding spatial details of 3D anatomical structures due to several key issues:

**Limited visual perception.** CLIP-style vision encoders [3, 14, 47, 45] are commonly utilized to extract discriminative visual features aligned with expert reports. However, unlike natural image-report datasets (e.g., 400M pairs [33]), the scarcity of 3D volume-report pairs (roughly 0.05M [14]) highly constrains feature space convergence. As a result, subtle but clinically critical pathological details may be obscured by irrelevant information, leading to suboptimal visual interpretation.

**Compromised semantic projection.** While multi-modal projectors aim to bridge vision and language by mapping 3D visual representations into LLM semantic spaces, current approaches (e.g., spatial pooling [3] and Q-former [25, 9]) struggle to preserve spatial and geometric details inherent in 3D anatomical structures. This limitation undermines the ability of LLMs to reason structural dependencies and pathological conditions, leading to unreliable outputs with fundamental errors.

In this paper, we propose Hybrid Spatial Encoding Network (HSENet), a novel framework that exploits enriched 3D medical visual cues with effective visual perception and projection for robust vision-language understanding. Specifically, to perceive spatial contexts from 3D volumetric space, we introduce a dual-stage 3D vision-language pretraining paradigm that trains dual-3D vision encoders: A 3D Vision Encoder learns global volumetric representations aligned with corresponding reports, while a 2D-Enhanced 3D Vision Encoder (2E3 Vision Encoder) refines report-aligned anatomical details, guided by the rich diagnostic insights recognized from 2D slices. Then, to map the extracted visual representations to LLM’s semantic space, we design Spatial Packer, an efficient projector that compresses 3D visual contexts into a compact set of informative visual tokens. This projector incorporates a novel Voxel2Point Cross-Attention (V2P-CA), which aggregates high-resolution 3D voxel representations to their centroid points, preserving essential spatial and geometric information. By integrating spatial packers with the pretrained dual-3D vision encoders, HSENet can effectively capture and transfer hybrid visual representations encompassing both global volumes and detailed anatomies, thereby enabling more accurate text generation. We provide comprehensive evaluations across 3D multi-modal retrieval, report generation, and VQA tasks. The results demonstrate that HSENet outperforms existing methods, achieving the state-of-the-art performance in generating discriminative visual representations and high-quality diagnostic responses.

## 2 Related Works

**Medical Multi-modal Large Language Models.** MLLMs have shown promise in vision-language applications within the medical field [19, 44]. Early explorations such as LLaVA-Med [24], Med-PaLM [39], Flamingo-CXR [37], and HuatuoGPT-Vision [7] integrate LLMs with 2D medical image encoders for diagnostic reasoning and achieve notable results. Building on this progress, RadFM [43], M3D [3], and CT-CHAT [14] extend MLLMs to 3D volumetric data, adapting them for various tasks, e.g., image-text retrieval, report generation, and VQA. However, these 3D foundational models rely on generic MLLM architectures that struggle to associate intricate 3D structures with medical language, resulting in hallucinations and factual errors. To address this, recent studies utilize advanced visual-language alignment strategies, including efficient pretraining [45, 4], knowledge injection [42, 21, 31, 4], and dedicated multi-modal projectors [36, 45, 11, 9]. Unlike the above methods, we introduce a hybrid visual perception and projection pipeline to distill enriched spatial patterns of global volume and local anatomy, enabling accurate and robust 3D vision-language understanding.

**3D Medical Vision-Language Alignment.** In medical MLLMs, learning aligned volume and report representations is essential for 3D downstream tasks [4]. Existing approaches can be broadly categorized into two stages: **1) Vision-language pre-training.** Xin et al. [45] leverages DCFormer [2] and pairwise sigmoid loss [46] to achieve efficient yet rich visual-textual alignment. Besides,

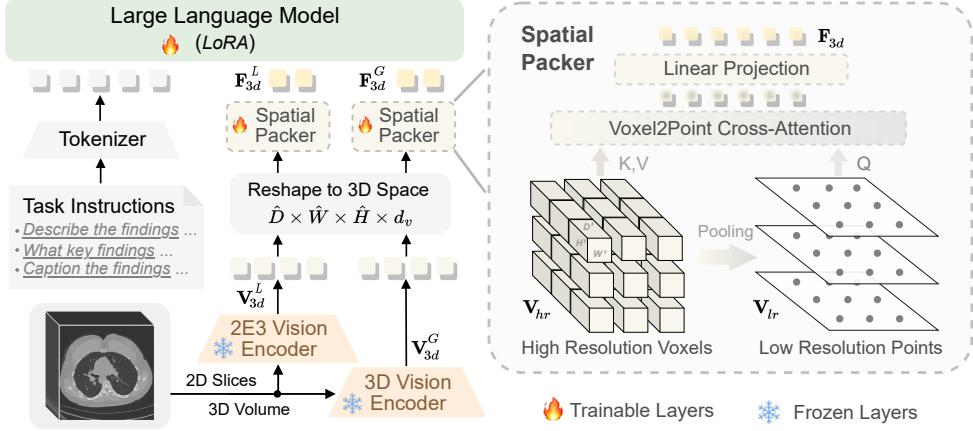


Figure 1: Architecture of the proposed HSENet. The input 3D CT volume is processed in parallel by the 3D Vision Encoder and the 2E3 Vision Encoder to extract rich, multi-scale features. These hybrid visual representations are then projected by two dedicated spatial packers into the semantic space of LLM, enabling effective 3D medical vision-language modeling.

additional supervision from external knowledge, such as electronic health records [4], medical entities [42], and LLM-summarized text [21] has also been shown to improve alignment quality. Nonetheless, abundant patient data is often difficult to obtain, while LLM-generated text may not always be reliable. In contrast, we leverage informative and readily accessible 2D slices from 3D volumes to promote vision-language consistency and strengthen 3D visual perception. **2) Multi-modal projection in MLLM fine-tuning.** Bai et al. [3] compress 3D tokens via spatial pooling to fit LLM input constraints, at the cost of losing spatial details. Med3DVLM [45] integrates MLP-Mixer [38] to capture hierarchical features and improve cross-modal interaction. Med-2E3 [36] projects both 2D slices and 3D volume features directly extracted from frozen encoders to the LLM, but may suffer from inconsistencies between 2D and 3D representations. In contrast, our approach decouples visual perception and projection processes. By utilizing spatial packers to independently project the visual contexts perceived by our pretrained, correlated dual visual encoders, we produce compact yet expressive hybrid representations that more effectively guide the LLM for clinical reasoning.

### 3 Methodology

#### 3.1 Overview

Given an input 3D CT volume  $\mathbf{I}_{3d} \in \mathbb{R}^{D \times W \times H \times C}$ , where  $D$ ,  $W$ ,  $H$ , and  $C$  represent the depth, width, height, and channel of the processed volume, respectively, our HSENet aims to learn rich visual representations and prompt the language model to generate the corresponding CT report  $R = \{r_1, \dots, r_M\}$  with  $M$  words. The architecture of HSENet is shown in Figure 1, which contains the encoding and projecting of hybrid visual features for accurate language generation.

**Hybrid Visual Encoding.** Clinically, the interpretation of 3D CT scans relies on both macro and micro levels of diagnosis, requiring observations of overall structures and detailed anatomical features [29]. Motivated by this, we introduce dual vision encoders to capture essential 3D medical information: a 3D Vision Encoder  $\mathbf{E}_{3d}(\cdot)$  for learning global volumetric structures, and a 2E3 Vision Encoder  $\mathbf{E}_{2e3}(\cdot)$  for learning local anatomical features. These encoders operate in parallel, extracting 3D features  $\mathbf{I}_{3d}$ , and generating global volumetric features  $\mathbf{V}_{3d}^G \in \mathbb{R}^{N_p \times d_v}$  and local anatomical features  $\mathbf{V}_{3d}^L \in \mathbb{R}^{N_p \times d_v}$ , respectively.  $N_p = (\hat{D} \times \hat{W} \times \hat{H})$  denotes the number of encoded 3D patches,  $(\hat{D}, \hat{W}, \hat{H})$  is the encoded spatial dimensions, and  $d_v$  is the feature dimension.

**Multi-modal Projection.** To effectively bridge the gap between 3D medical images and the LLM’s semantic space, we introduce a spatial packer that condenses high-resolution 3D regions into a compact set of visual tokens. Specifically, we employ twin spatial packers to process global ( $\mathbf{V}_{3d}^G$ )

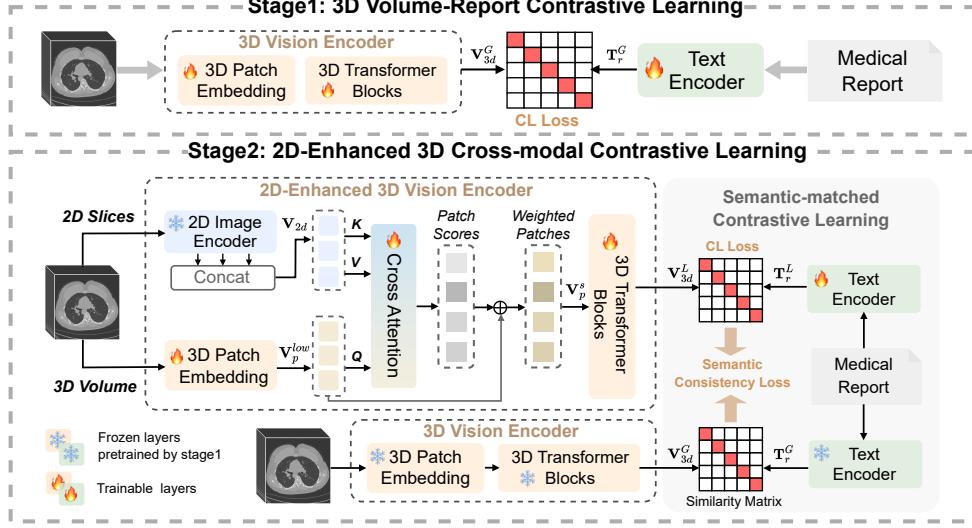


Figure 2: Overview of the dual-stage pretraining framework. **Stage 1:** The 3D Vision Encoder is trained for global vision-language alignment using paired 3D volumes and medical reports. **Stage 2:** The 2E3 Vision Encoder is trained to exploit anatomy-related local 3D patches aligned with reports. A semantic consistency loss is applied in Stage 2 to maintain alignment with the global relations learned in Stage 1, ensuring a stable local representation refining.

and local ( $\mathbf{V}_{3d}^L$ ) 3D visual features in parallel, resulting in transformed features  $\mathbf{F}_{3d}^G \in \mathbb{R}^{N'_p \times d_t}$  and  $\mathbf{F}_{3d}^L \in \mathbb{R}^{N'_p \times d_t}$ . Here,  $N'_p$  denotes the number of compressed tokens,  $d_t$  is LLM’s feature dimension.

**Language Decoding.** We construct multi-modal prompts by concatenating the projected hybrid visual representations with task instructions, guiding the LLM to generate diagnostic answers. To optimize the LLM, we employ LoRA [15] and minimize the following cross-entropy loss:

$$\mathcal{L}_{Gen} = - \sum_{t=1}^M \log P(y_t | y_{1:t-1}, \{\mathbf{F}_{3d}^G, \mathbf{F}_{3d}^L\}; \theta), \quad (1)$$

where  $P(y_t | *)$  denotes the probability of predicting text token  $y_t$  conditioned on the preceding tokens  $y_{1:t-1}$  and the projected hybrid visual features  $\mathbf{F}_{3d}^G$  and  $\mathbf{F}_{3d}^L$ .  $\theta$  denotes the trainable parameters.

### 3.2 Dual-stage 3D Medical Vision-Language Pretraining

To mimic the way physicians observe macro and micro 3D visual patterns, we design a novel dual-stage cross-modal pretraining framework to build robust 3D vision encoders. As illustrated in Figure 2, we first conduct 3D volume-report contrastive learning to train a 3D Vision Encoder for capturing macro-level CT structures. Then, we propose 2D-enhanced 3D (2E3) cross-modal contrastive learning to refine the 2E3 Vision Encoder by incorporating detailed 3D anatomical patterns, leveraging cross-modal relations, enriched semantics from related 2D slices.

**Stage 1: 3D Volume-Report Contrastive Learning.** We harness expert-written reports as inherent labels to learn discriminative visual representations of 3D CT volumes. Following common paradigms [33, 47], we pair a 3D vision encoder  $E_{3d}(\cdot)$  and a text encoder  $E_{text}^{s1}(\cdot)$  to extract volume features  $\mathbf{V}_{3d}^G$  and report features  $\mathbf{T}_r^G$ , respectively. To align these features, we leverage the CLS token from each encoder as a compact summary embedding, which is then projected into a shared latent space  $\tilde{\mathbf{x}}_{3d} \in \mathbb{R}^{d_t}$  and  $\tilde{\mathbf{x}}_t \in \mathbb{R}^{d_t}$ . The objective of this stage is to maximize the mutual information between paired volume and report, achieved by optimizing symmetric InfoNCE [40] loss:

$$\mathcal{L}_{CL} = -\frac{1}{2N_c} \sum_{i=1}^{N_c} \left( \log \frac{\exp(\text{sim}(\tilde{\mathbf{x}}_{3d}^{(i)}, \tilde{\mathbf{x}}_t^{(i)})/\tau)}{\sum_{k=1}^B \exp(\text{sim}(\tilde{\mathbf{x}}_{3d}^{(i)}, \tilde{\mathbf{x}}_t^{(k)})/\tau)} + \log \frac{\exp(\text{sim}(\tilde{\mathbf{x}}_t^{(i)}, \tilde{\mathbf{x}}_{3d}^{(i)})/\tau)}{\sum_{k=1}^B \exp(\text{sim}(\tilde{\mathbf{x}}_t^{(i)}, \tilde{\mathbf{x}}_{3d}^{(k)})/\tau)} \right), \quad (2)$$

where  $\text{sim}(\cdot)$  computes the cosine similarity,  $B$  is the batch size,  $\tau$  is the temperature hyperparameter, and  $N_c$  denotes the number of volume-report pairs.

**Stage 2: 2D-Enhanced 3D Cross-modal Contrastive Learning.** Radiologists are skilled in correlating 3D contextual information with 2D slice-level observations to interpret subtle anatomies [34]. Inspired by this, we distill knowledge in 2D slices to refine 3D vision-language alignment from global to fine-grained anatomy. This approach is more promising than current methods that rely on external patient data [42], which is often inaccessible, or on potentially unreliable LLM-generated texts [21], since our 2D slices can be readily obtained from 3D volumes and inherently contain rich diagnostic information.

**2D Slice Processing.** We uniformly slice the 3D volume along the Z-axis and obtain  $\mathbf{I}_{2d} = \{s_1, s_2, \dots, s_{N_s}\}$ , where  $N_s$  represents the number of extracted slices. We extract slice features by processing each slice with pre-trained BioMedCLIP [47], then stacking them into  $\mathbf{V}_{2d} \in \mathbb{R}^{N_s \times d_v}$ .

**2D-Enhanced 3D Vision Enhancing.** Unlike previous methods that focus on augmenting high-level 3D visual features [21, 42], we argue that low-level features carry richer 3D spatial cues for capturing anatomical details and improving visual representation quality. Accordingly, as illustrated in the bottom of Figure 2, we introduce a 2D-enhanced 3D vision encoder  $\mathbf{E}_{2e3}(\cdot)$  for local vision enhancement. Firstly, we extract low-level 3D patch features  $\mathbf{V}_p^{low} \in \mathbb{R}^{N_p \times d_v}$  from the 3D patch embedding layer of a standard 3D ViT. We then interact  $\mathbf{V}_p^{low}$  with 2D features  $\mathbf{V}_{2d}$  by cross-attention layers, to estimate the significance of distinct 3D patches:

$$\mathbf{S}_{3d} = \text{FFN}(\text{MHA}(\mathbf{V}_p^{low}, \mathbf{V}_{2d}, \mathbf{V}_{2d})), \quad (3)$$

where  $\text{FFN}$  and  $\text{MHA}$  denote feed-forward and multi-head attention layers. The resulting scoring feature  $\mathbf{S}_{3d} \in \mathbb{R}^{N_p \times d}$  is then projected via MLP layers to produce patch scores  $\mathbf{S}'_{3d} = \{s_{3d}^{(1)}, s_{3d}^{(2)}, \dots, s_{3d}^{(N_p)}\} \in \mathbb{R}^{N_p}$ , with  $s_{3d}^{(i)}$  indicating the importance of the  $i$ -th 3D patch. Using these scores, we weight the low-level 3D patch features, yielding  $\mathbf{V}_p^s \in \mathbb{R}^{N_p \times d_v}$ , which emphasizes diagnostically relevant spatial areas. Finally,  $\mathbf{V}_p^s$  is fed through transformer blocks to generate high-level vision features  $\mathbf{V}_{3d}^L$  that capture local 3D anatomical details.

**Semantic-Matched Contrastive Learning.** To capture fine-grained anatomical representations, we apply contrastive learning loss  $\mathcal{L}_{CL}^{2e3}$  similar to Equation 2, aligning the enhanced 3D visual features  $\mathbf{V}_{3d}^L$  with the corresponding report features produced by text decoder  $\mathbf{E}_{text}^{s_2}(\cdot)$ . While this objective encourages detailed local alignment, unconstrained optimization risks drifting from generalizable vision-text relationships. To mitigate this, we introduce a semantic consistency loss  $\mathcal{L}_{SA}$  that regularizes the cross-modal similarity matrix by anchoring it to the global alignment established in Stage 1. The loss function is formulated as:

$$\mathcal{L}_{SA} = \sum_{i=1}^B \left\| \text{sim}(\tilde{\mathbf{x}}_{3d}^{(i)}, \tilde{\mathbf{x}}_{t1}^{(i)})/\tau - \text{sim}(\tilde{\mathbf{x}}_{2e3}^{(i)}, \tilde{\mathbf{x}}_{t2}^{(i)})/\tau \right\|^2, \quad (4)$$

where  $\tilde{\mathbf{x}}_{3d}$ ,  $\tilde{\mathbf{x}}_{t1}$  are the volume and report features from the fixed Stage 1 encoders  $\mathbf{E}_{3d}(\cdot)$  and  $\mathbf{E}_{text}^{s_1}(\cdot)$ , respectively.  $\tilde{\mathbf{x}}_{2e3}$  and  $\tilde{\mathbf{x}}_{t2}$  are the local vision features and report features from stage 2. The overall loss in stage 2 can be calculated as:

$$\mathcal{L}_{SCL} = \mathcal{L}_{CL}^{2e3} + \lambda_s \mathcal{L}_{SA}, \quad (5)$$

where  $\lambda_s$  controls the regularization strength. During Stage 2, the Stage 1 encoders are frozen, while  $\mathbf{E}_{2e3}(\cdot)$  and  $\mathbf{E}_{text}^{s_2}(\cdot)$  are trainable. This formulation preserves foundational knowledge from Stage 1 while refining representations in Stage 2, enhancing the model’s capacity to capture fine-grained anatomical details and maintain robust vision-text alignment.

### 3.3 Spatial Packer

As shown in Figure 1, we propose spatial packers to project the extracted global and local 3D visual features ( $\mathbf{V}_{3d}^G$  and  $\mathbf{V}_{3d}^L$ ) into LLM’s latent space. The key insight behind spatial packer is to leverage both high- and low-resolution embeddings for efficient token compression and spatial preservation. Here, we illustrate the workflow of spatial packer using  $\mathbf{V}_{3d}^G$  as a representative example.

**High-Resolution Voxel Embedding.** Following Bai et al. [3], we reshape the patch dimension  $N_p$  of  $\mathbf{V}_{3d}^G \in \mathbb{R}^{N_p \times d_v}$  back to its original 3D spatial layout, obtaining  $\mathbf{V}_{3d'}^G \in \mathbb{R}^{\hat{D} \times \hat{W} \times \hat{H} \times d_v}$ . We then

partition  $\mathbf{V}_{3d'}^G$  along each spatial axis using strides  $(S_d, S_w, S_h)$ , resulting in high-resolution voxel features  $\mathbf{V}_{hr}^G \in \mathbb{R}^{(S_d \cdot S_w \cdot S_h) \times D' \times W' \times H' \times d_v}$ , where  $D' = \frac{\hat{D}}{S_d}$ ,  $W' = \frac{\hat{W}}{S_w}$ , and  $H' = \frac{\hat{H}}{S_h}$  denotes the spatial dimensions of each local voxel (see right part of Figure 1).  $\mathbf{V}_{hr_{i,j,k}}^G \in \mathbb{R}^{D' \times W' \times H' \times d_v}$  represents the spatial feature of the voxel coordinated at  $(i, j, k)$  in volume space.

**Low-Resolution Point Embedding.** To capture the overall pattern within each local voxel, we apply feature pooling for  $\mathbf{V}_{hr_{i,j,k}}^G \in \mathbb{R}^{D' \times W' \times H' \times d_v}$ , extracting a centroid point representation  $\mathbf{V}_{lr_{i,j,k}}^G \in \mathbb{R}^{d_v}$ . For the entire 3D volume, these centroid embeddings aggregated into the low-resolution point embedding  $\mathbf{V}_{lr}^G \in \mathbb{R}^{S_d \times S_w \times S_h \times d_v}$ , where  $S_d$ ,  $S_w$ , and  $S_h$  denote the number of points along each spatial dimension.

**Voxel2Point Cross-Attention.** We propose a Voxel2Point Cross-Attention (V2P-CA) mechanism to inject enriched spatial clues from high-resolution  $\mathbf{V}_{hr}^G$  into low-dimensional  $\mathbf{V}_{lr}^G$ , enabling efficient visual projection. Unlike previous cross-attention-based projectors [28, 27, 8] that are limited to 2D images, our V2P-CA learns 3D voxel-point interactions for effective spatial preservation. We first reshape  $\mathbf{V}_{lr}^G$  as low-resolution query  $Q_l \in \mathbb{R}^{(S_d \cdot S_w \cdot S_h) \times 1 \times d_v}$ , and reshape  $\mathbf{V}_{hr}^G$  as high-resolution key  $K_h \in \mathbb{R}^{(S_d \cdot S_w \cdot S_h) \times (D' \cdot W' \cdot H') \times d_v}$  and value  $V_h \in \mathbb{R}^{(S_d \cdot S_w \cdot S_h) \times (D' \cdot W' \cdot H') \times d_v}$ . Then, we leverage cross-attention to make each point in  $Q_l$  fully absorb its corresponding fine-grained voxel features in  $K_h$  and  $V_h$ :

$$\mathbf{Y}_{3d}^G = FFN(MHA(Q_l, K_h, V_h)), \quad (6)$$

where  $\mathbf{Y}_{3d}^G \in \mathbb{R}^{(S_d \cdot S_w \cdot S_h) \times d_v}$  denotes the compact spatial visual tokens. We finally use 2-layer MLPs to map the  $\mathbf{Y}_{3d}^G$  to LLM’s latent dimension, producing  $\mathbf{F}_{3d}^G \in \mathbb{R}^{N'_p \times d_t}$  ( $N'_p = S_d \cdot S_w \cdot S_h$ ). We adopt the same procedure to generate  $\mathbf{F}_{3d}^L$  for local anatomical features  $\mathbf{V}_{3d}^L$ .

## 4 Experiments and Results

### 4.1 Experiment Settings

**Tasks and Datasets.** To validate HSENet for 3D medical vision-language understanding, we evaluate on three tasks: (1) medical image-text retrieval, (2) report generation, and (3) medical VQA. For image-text retrieval and report generation tasks, we use the benchmark 3D CT dataset CT-RATE [14], which contains 25,692 non-contrast chest CT scans from 21,304 anonymized patients. After data expansion and excluding cases with excessively short or invalid reports, we retain 47,149 volume-report pairs (20,000 unique patients) for training and 3,039 pairs (1,304 distinct patients) for testing. For medical VQA, we adopt the RadGenome-ChestCT dataset [50], which contains 302,827 open-ended VQA pairs focused on 3D location observations, enabling the evaluation of models’ 3D spatial reasoning capabilities. We allocate 285,086 samples for training and 17,741 samples for testing.

**Implementation Details.** We employ the standard 3D Vision Transformer (3D ViT) [13] and Bert [12] as visual and language encoders for pretraining and retrieval. We utilize Phi4-4B-Instruct [1] as the language model, which is integrated with our pretrained dual visual encoders and spatial packers to construct MLLM. Following Bai et al. [3], input volumes are normalized and resized to  $(D, W, H) = (32, 256, 256)$  using Min-Max Normalization, then encoded into patches of size  $(\hat{D}, \hat{W}, \hat{H}) = (8, 16, 16)$  by 3D ViT. The spatial packer uses strides  $(S_d, S_w, S_h) = (8, 4, 4)$ , yielding local voxels of size  $(D', W', H') = (1, 4, 4)$ . We set the number of 2D slices  $N_s = 32$ , loss weight  $\lambda_s = 0.1$ , and feature dimensions  $d_v = 768$ ,  $d_l = 512$ ,  $d_t = 3072$ . Experiments are conducted on 8 RTX 3090 GPUs using AdamW optimizer. Both pretraining stages run for 50 epochs with a learning rate of 1e-4. Report generation is trained for 6 epochs at 1e-4 and VQA for 4 epochs at 5e-5. Additional details are provided in the supplementary material.

**Evaluation Metrics.** We use Recall@K (R@5/10/50/100) to evaluate top-k retrieval accuracy in report-to-volume and volume-to-report tasks. For volume-to-volume retrieval, we utilize Mean Average Precision (MAP@5/10/50) to assess the model’s ability to retrieve pathology-relevant volumes. Report generation is evaluated using standard natural language generation (NLG) metrics (BLEU [32], ROUGE [30], METEOR [22], and BERTScore [48]) to measure linguistic quality, along with RaTE-Score [51] to assess clinical relevance. For the VQA task, we use both the NLG metrics and answer accuracy. The accuracy evaluates the performance separately on major (e.g., lung, heart) and minor (e.g., left lung lower lobe, left heart ventricle) location categories.

Table 1: Experiments on image-text retrieval performance. **Bold** indicates the best performance, while underlined indicates the second-best performance for each model. *3D-ViT* and *2E3-ViT* refer to our 3D Vision Encoder  $\mathbf{E}_{3d}(\cdot)$  and 2E3 Vision Encoder  $\mathbf{E}_{2e3}(\cdot)$ , respectively.  $\dagger$  denotes the model reproduced using the official code. *Full Text*, *Text CLS*, and *2D Slices* refer to the features used for guiding patch scoring within  $\mathbf{E}_{2e3}(\cdot)$ .

Methods	Report-to-Volume Retrieval				Volume-to-Report Retrieval				Volume-to-Volume Retrieval		
	R@5	R@10	R@50	R@100	R@5	R@10	R@50	R@100	MAP@5	MAP@10	MAP@50
<i>(a) comparison with state-of-the-art pretraining models</i>											
VocabFine[14]	0.10	0.60	2.30	2.00	/	/	/	/	68.30	57.20	48.80
MG-3D[31]	/	/	3.88	/	/	/	/	/	/	/	/
Merlin[4]	1.50	2.70	7.70	12.70	/	/	/	/	62.60	51.30	43.90
CT-CLIP[14]	2.90	5.00	18.00	28.70	/	/	/	/	68.30	57.20	48.90
M3D-CLIP[3] $\dagger$	4.87	8.72	24.42	33.89	5.30	8.88	24.38	34.16	68.80	57.83	49.54
Med3DVLM[45] $\dagger$	2.96	4.94	15.56	23.89	2.44	3.78	12.41	18.33	68.31	56.98	48.31
Ours (3D-ViT)	<b>5.76</b>	<b>9.28</b>	<u>25.50</u>	<u>34.72</u>	<b>5.63</b>	<b>9.05</b>	<b>25.67</b>	<b>34.62</b>	68.75	<b>57.85</b>	<b>49.57</b>
Ours (2E3-ViT)	<b>5.82</b>	<b>9.44</b>	<b>28.46</b>	<b>39.85</b>	<b>6.09</b>	<b>9.67</b>	<b>28.63</b>	<b>39.22</b>	<b>69.32</b>	<b>58.68</b>	<b>50.58</b>
<i>(b) different settings for 3D patch scoring</i>											
Full Text	1.61	3.26	10.89	16.72	1.51	2.96	10.53	16.42	66.56	55.17	46.93
Text CLS	2.93	<u>5.76</u>	<u>19.32</u>	29.48	3.32	6.12	19.78	28.59	68.00	57.10	48.85
2D Slice	<b>5.82</b>	<b>9.44</b>	<b>28.46</b>	<b>39.85</b>	<b>6.09</b>	<b>9.67</b>	<b>28.63</b>	<b>39.22</b>	<b>69.32</b>	<b>58.68</b>	<b>50.58</b>
<i>(c) ablation study of semantic consistency loss</i>											
w/o $\mathcal{L}_{SA}$	4.90	8.29	<u>27.28</u>	37.97	4.77	9.15	26.82	37.94	69.12	58.45	50.38
Ours (2E3-ViT)	<b>5.82</b>	<b>9.44</b>	<b>28.46</b>	<b>39.85</b>	<b>6.09</b>	<b>9.67</b>	<b>28.63</b>	<b>39.22</b>	<b>69.32</b>	<b>58.68</b>	<b>50.58</b>

## 4.2 Results on Medical Image-Text Retrieval

To assess the capability of the pretrained dual 3D vision encoders, we evaluate their effectiveness across various retrieval tasks: 1) report-to-volume retrieval, 2) volume-to-report retrieval, and 3) volume-to-volume retrieval.

### Comparisons with 3D Pretraining Models.

We compare the performance with state-of-the-art 3D medical vision-language pretraining models, including CT-CLIP [14], M3D-CLIP [3], VocabFine [14], MG-3D [31], Merlin [4], and Med3DVLM [45]. As evidenced by Table 1(a), our method achieves consistent superiority across all evaluation metrics in three retrieval tasks. Notably, our 3D Vision Encoder, based on simple 3D patch processing, achieves  $1.99\times$  higher R@5 in report-to-volume retrieval compared to the more complex hierarchical volume partitioning and encoding pipeline in CT-CLIP (5.76% vs. 2.90%). By incorporating 2D slice-guided patch scoring, our 2E3 Visual Encoder yields substantial gains, improving R@100 by  $\sim 5\%$  over the 3D Vision Encoder. This indicates that learning local anatomical patterns from 2D slices can effectively model the intricate 3D volume-report relations. The 2E3 Vision Encoder also achieves SoTA volume-to-volume retrieval performance, suggesting that our model learns discriminative 3D medical features, thereby retrieving volumes with high pathological relevance.

**Ablation Studies.** To evaluate the effectiveness of our 2D-guided patch scoring, we replaced 2D slice features with alternative text-derived features, including the full text embedding  $\mathbf{T}_{full} \in \mathbb{R}^{512 \times 768}$  and the CLS token  $\mathbf{T}_{cls} \in \mathbb{R}^{768}$  from the text encoder  $\mathbf{E}_{text}^{s2}(\cdot)$ . As shown in Table 1(b), both variants lead to performance drops, particularly with  $\mathbf{T}_{full}$  (R@5: from 5.82% to 1.61%, 4.21% $\downarrow$ ). This

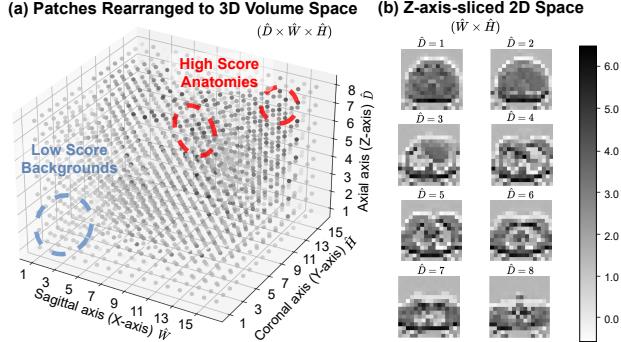


Figure 3: Visualization of 3D patch scores in 2E3 Vision Encoder. Darker colors indicate higher scores. (a) Patches are rearranged into the original 3D volume space to illustrate their score distribution. The model assigns higher scores to semantically essential patches (highlighted in red) and lower scores to less relevant patches (in blue). (b) Axial slices along the Z-axis reveal the patch scores at different depth levels  $\hat{D}$ , providing a clearer view of the score variations.

Table 2: Experiments on report generation. **Bold** indicates the best performance, while underlined indicates the second-best performance.  $\dagger$  denotes the model reproduced using the official code.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-L	METEOR	BERT-Score	RaTE-Score
<i>(a) comparison with state-of-the-art models</i>									
RadFM[43]	29.85	/	/	/	45.67	/	28.75	86.97	/
CT-CHAT[14]	39.52	/	/	/	/	32.12	21.85	/	/
M3D-LaMed[3]	40.32	/	/	/	52.08	/	36.67	87.55	/
E3D-GPT[20]	41.15	/	/	/	52.60	/	41.79	87.97	/
Med-2E3[36] $\dagger$	55.87	30.82	19.64	14.09	54.40	33.33	43.06	87.99	<u>61.81</u>
Med3DVLM[45] $\dagger$	<u>56.76</u>	<u>32.20</u>	21.46	<u>16.00</u>	54.38	<u>34.17</u>	<u>43.18</u>	<u>88.12</u>	<u>61.07</u>
HSENet (Ours)	<b>62.89</b>	<b>39.47</b>	<b>29.11</b>	<b>24.01</b>	<b>56.50</b>	<b>40.63</b>	<b>44.75</b>	<b>88.99</b>	<b>64.99</b>
<i>(b) comparison of different multi-modal projectors utilized in HSENet</i>									
Q-Former[25]	55.60	32.30	22.15	17.11	53.62	35.47	43.29	87.97	62.39
Sequence Pooling[3]	56.20	33.40	23.40	18.46	53.61	36.29	43.51	88.08	62.82
Spatial Pooling[3]	<u>61.67</u>	<u>37.20</u>	<u>26.14</u>	<u>20.66</u>	<u>56.48</u>	<u>38.54</u>	<u>44.21</u>	<u>88.84</u>	<u>63.16</u>
Spatial Packer	<b>62.89</b>	<b>39.47</b>	<b>29.11</b>	<b>24.01</b>	<b>56.50</b>	<b>40.63</b>	<b>44.75</b>	<b>88.99</b>	<b>64.99</b>

Table 3: Ablation studies on different settings of the visual encoders and the projector in medical report generation. *3D-ViT* and *2E3-ViT* denotes the proposed 3D Visual Encoder  $E_{3d}(\cdot)$  and 2E3 Visual Encoder  $E_{2e3}(\cdot)$ , respectively.

Methods	Vision Encoder	Spatial Packer	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-L	METEOR	BERT-Score	RaTE-Score	
	3D-ViT	2E3-ViT										
(a)	$\checkmark$		55.57	32.81	22.88	17.93	53.16	35.86	43.14	87.96	62.35	
(b)	$\checkmark$	$\checkmark$	58.69	34.24	23.43	17.87	55.06	35.84	43.59	88.34	62.74	
(c)		$\checkmark$	57.87	33.81	23.17	17.79	54.86	35.92	43.69	88.25	62.84	
(d)	$\checkmark$	$\checkmark$	60.96	36.37	25.44	19.95	56.06	37.65	43.99	88.69	<u>63.37</u>	
(e)	$\checkmark$	$\checkmark$	<u>61.67</u>	<u>37.20</u>	<u>26.14</u>	<u>20.66</u>	<u>56.48</u>	<u>38.54</u>	<u>44.21</u>	<u>88.84</u>	<u>63.16</u>	
Ours	$\checkmark$	$\checkmark$	$\checkmark$	<b>62.89</b>	<b>39.47</b>	<b>29.11</b>	<b>24.01</b>	<b>56.50</b>	<b>40.63</b>	<b>44.75</b>	<b>88.99</b>	<b>64.99</b>

may be due to the inherent gap between high-level textual features and our low-level visual patches, which prevents visual enhancement as achieved by 2D slices. Additionally, removing the semantic consistency loss  $\mathcal{L}_{SA}$  also results in performance degradation (see Table 1(c)). This confirms its importance in maintaining stable cross-modal correspondence during local feature refinement.

### Visualization of 3D

**Patch Scores.** Figure 3 visualizes the 3D patch scores produced by the 2E3 Vision Encoder, demonstrating its capacity to differentiate informative regions from irrelevant ones. Both volumetric (3D) and slice-based (2D) views are presented to show the spatial distribution of scores. The model consistently assigns higher scores to anatomically salient patches, thereby enhancing the effectiveness of local representation learning.

Table 4: Experiments on 3D medical VQA. *Major Class Acc* measures the accuracy in answering the major location category, while *Minor Class Acc* evaluates accuracy on more detailed body locations.

Methods	BLEU-1	ROUGE-1	METEOR	BERT-Score	Major Class Acc.	Minor Class Acc.
M3D-LaMed[3] $\dagger$	60.15	56.49	21.25	90.36	71.61	28.08
Med-2E3[36] $\dagger$	63.45	52.36	17.44	89.84	66.70	25.71
Med3DVLM[45] $\dagger$	<u>64.11</u>	57.08	19.59	90.65	70.39	28.90
Ours (3D-ViT)	<b>59.70</b>	<b>56.80</b>	<b>21.66</b>	<b>90.56</b>	<b>71.07</b>	<b>28.84</b>
Ours (2E3-ViT)	61.17	<u>57.85</u>	<u>21.74</u>	<u>90.71</u>	<u>72.28</u>	<u>29.59</u>
Ours (Dual-ViT)	<b>65.65</b>	<b>58.90</b>	<b>21.58</b>	<b>90.77</b>	<b>73.60</b>	<b>30.28</b>

### 4.3 Results on Medical Report Generation

**Comparison Studies.** We compare the report generation performance with advanced 3D medical MLLMs, including RadFM [43], CT-CHAT [14], M3D-LaMed [3], E3D-GPT [20], Med-2E3 [36], and Med3DVLM [45]. As shown in Table 2(a), our model achieves state-of-the-art performance across all NLG metrics and the RaTE-Score, reflecting both linguistic fluency and clinical accuracy. Foundation models such as RadFM, CT-CHAT, and M3D-LaMed mainly adopt generic MLLM architectures and lack dedicated designs to grasp 3D medical clues, leading to lower overall scores. Med3DVLM uses DCFormer for multi-scale volumetric features and improves BLEU scores, while Med-2E3 enhances clinical relevance (RaTE-Score +0.74%) by fusing 2D and 3D features for LLM inference, though sacrificing coherence (only 14.09% of BLEU-4). In contrast, our method effectively decouples the 3D perception and projection, yielding superior overall results.

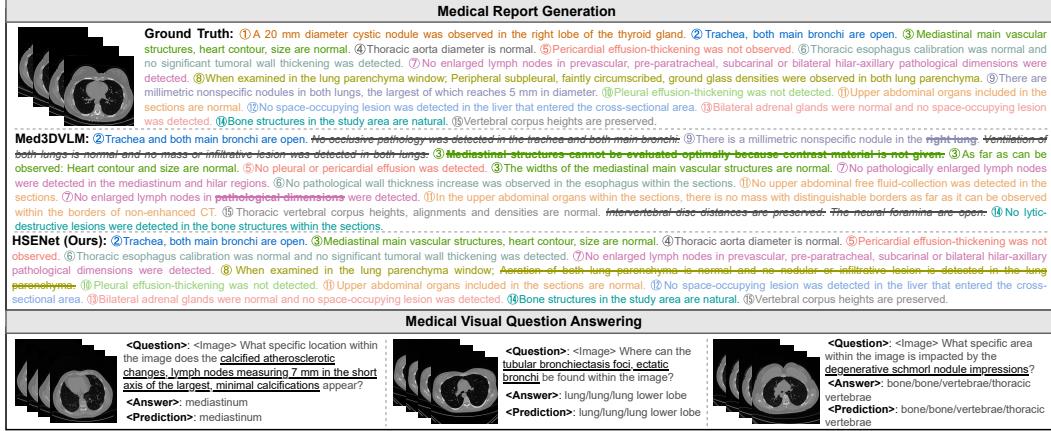


Figure 4: Visualization of 3D CT report generation and medical VQA. Different colors in reports highlight distinct diagnostic findings. Strikethrough marks incorrect predictions, while *italicized words* indicate generated contents absent from ground truth reports.

**Ablation Studies.** We perform ablation studies to assess the impact of the spatial packer and dual visual encoders. As shown in Table 2, replacing our spatial packer with Q-Former or pooling strategies degrades performance, with Q-Former leading to a 7.29% BLEU-1 drop, likely due to disrupted 3D structure. Table 3 compares different visual encoder configurations: Results from settings (a), (c), and (e) show that the 2E3 Vision Encoder outperforms 3D Vision Encoder (+2.3% BLEU-1), and combining both encoders further improves performance by using complementary hybrid 3D features.

**Visualization of Report Generation.** Figure 4 visualizes reports generated by our model and the most advanced Med3DVLM. We find that Med3DVLM exhibits notable errors and hallucinations in diagnosing 3D organs, highlighting the challenges of understanding 3D spatial patterns. In contrast, our HSENet produces more accurate diagnoses and identifies key structures, such as “*bilateral adrenal glands*” and “*thoracic aorta*”, which Med3DVLM overlooks. These results further demonstrate the strength of our hybrid visual contexts in capturing 3D spatial information.

#### 4.4 Results on Medical VQA

We also assess the model’s spatial reasoning capability via medical visual question answering. As shown in Table 4, our approach surpasses prior methods, with accuracy gains of +3.21% (major classes) and +1.38% (minor classes) over Med3DVLM. Notably, we find that using only the 2E3 Visual Encoder already yields notable gains over the 3D Visual Encoder competitor, achieving 1.47%↑ on BLEU-1 and 1.21%↑ on Major Class Accuracy. This suggests that our model relies more on the local 3D features given by 2E3 Visual Encoder for reasoning spatial locations. By aggregating both the local and global 3D representations, our HSENet captures richer visual contexts and achieves the best performance, which is aligned with the findings of Huang et al. [17]. Qualitative results in Figure 4 further demonstrate HSENet’s ability to infer precise 3D locations in VQA scenarios.

## 5 Conclusion

We present HSENet, a novel 3D medical vision-language model that bridges the visual perception and projection to understand complex 3D spatial structures for CT diagnosis. HSENet introduces dual 3D vision encoders to perceive both global volumetric context and local anatomical details, and designs a spatial packer to project 3D spatial features into the LLM’s semantic space via compact, informative tokens. We conduct comprehensive evaluations on benchmark datasets across 3D multi-modal retrieval, report generation, and medical VQA tasks. HSENet achieves state-of-the-art performance in both visual representation learning and diagnostic text generation. We believe this work can provide promising insights toward unified 3D image-report understanding and inspire further research in enhancing computer-aided CT diagnosis.

## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61906007, 62276010, and 62306253, in part by the Guangdong Natural Science Fund-General Programme under Grant 2024A1515010233.

## References

- [1] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyra Zhang, Yunan Zhang, und Xiren Zhou. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. [CoRR](#), abs/2503.01743, 2025.
- [2] Gorkem Can Ates, Kuang Gong, und Wei Shao. DCFormer: Efficient 3D Vision-Language Modeling with Decomposed Convolutions. [CoRR](#), abs/2502.05091, 2025.
- [3] Fan Bai, Yuxin Du, Tiejun Huang, Max Qinghu Meng, und Bo Zhao. M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models. [CoRR](#), abs/2404.00578, 2024.
- [4] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Pontes Reis, Cesar Truyts, Christian Bluethgen, Malte Engmann Kjeldskov Jensen, Sophie Ostmeier, Maya Varma, Jeya Maria Jose Valanarasu, Zhongnan Fang, Zepeng Huo, Zaid Nabulsi, Diego Ardila, Wei-Hung Weng, Edson Amaro Júnior, Neera Ahuja, Jason Alan Fries, Nigam H. Shah, Andrew Johnston, Robert D. Boutin, Andrew Wentland, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, und Akshay S. Chaudhari. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. [CoRR](#), abs/2406.06512, 2024.
- [5] Michael A. Bruno, Eric A. Walker, und Hani H. Abujudeh. Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. *Radiographics : a review publication of the Radiological Society of North America, Inc*, 35 6 S. 1668–76, 2015.
- [6] Hao Chen, Wei Zhao, Yingli Li, Tianyang Zhong, Yisong Wang, Youlan Shang, Lei Guo, Junwei Han, Tianming Liu, Jun Liu, und Tuo Zhang. 3D-CT-GPT: Generating 3D Radiology Reports through Integration of Large Vision-Language Models. [CoRR](#), abs/2409.19330, 2024.
- [7] Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, und Benyou Wang. HuatuoGPT-Vision, Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. [CoRR](#), abs/2406.19280, 2024.
- [8] Kezhen Chen, Rahul Thapa, Rahul Chalamala, Ben Athiwaratkun, Shuaiwen Leon Song, und James Zou. Dragonfly: Multi-Resolution Zoom Supercharges Large Visual-Language Model. [CoRR](#), abs/2406.00977, 2024.
- [9] Qiuwei Chen, Xinyue Hu, Zirui Wang, und Yi Hong. MedBLIP: Bootstrapping Language-Image Pre-training from 3D Medical Images and Texts. [CoRR](#), abs/2305.10799, 2023.
- [10] Yinda Chen, Che Liu, Xiaoyu Liu, Rossella Arcucci, und Zhiwei Xiong. BIMCV-R: A Landmark Dataset for 3D CT Text-Image Retrieval. In *Medical Image Computing and Computer Assisted Intervention, MICCAI 2024*, volume 15011 of *Lecture Notes in Computer Science*, S. 124–134, 2024.

- [11] Zhixuan Chen, Luyang Luo, Yequan Bie, und Hao Chen. Dia-LLaMA: Towards Large Language Model-driven CT Report Generation. [CoRR](#), abs/2403.16386, 2024.
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, und Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In [Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies](#), NAACL-HLT 2019, S. 4171–4186, 2019.
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, und Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In [9th International Conference on Learning Representations](#), ICLR 2021, 2021.
- [14] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Kumar Sekuboyina, Berkcan Lafci, Mehmet Kemal Ozdemir, und Bjoern H Menze. Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography. [arXiv preprint arXiv:2403.17834](#), 2024.
- [15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, und Weizhu Chen. Lora: Low-rank adaptation of large language models. [arXiv preprint arXiv:2106.09685](#), 2021.
- [16] Shih-Cheng Huang, Tanay Kothari, Imon Banerjee, Christopher Chute, Robyn L. Ball, Norah Borus, Andrew Huang, Bhavik N. Patel, Pranav Rajpurkar, Jeremy Irvin, Jared Dunnmon, Joseph Bledsoe, Katie S. Shpanskaya, Abhay Dhaliwal, Roham Zamani, Andrew Y. Ng, und Matthew P. Lungren. PENet - a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. [npj Digit. Medicine](#), 3, 2020.
- [17] Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, und Serena Yeung. GLoRIA: A Multi-modal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition. In [2021 IEEE/CVF International Conference on Computer Vision](#), ICCV 2021, S. 3922–3931, 2021.
- [18] Baoyu Jing, Pengtao Xie, und Eric P. Xing. On the Automatic Generation of Medical Imaging Reports. In [Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics](#), ACL 2018, S. 2577–2586, 2018.
- [19] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, und Hae Won Park. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. In [Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024](#), NeurIPS 2024, 2024.
- [20] Haoran Lai, Zihang Jiang, Qingsong Yao, Rongsheng Wang, Zhiyang He, Xiaodong Tao, Wei Wei, Weifu Lv, und S. Kevin Zhou. E3D-GPT: Enhanced 3D Visual Foundation for Medical Vision-Language Model. [CoRR](#), abs/2410.14200, 2024.
- [21] Haoran Lai, Zihang Jiang, Qingsong Yao, Rongsheng Wang, Zhiyang He, Xiaodong Tao, Wei Wei, Weifu Lv, und S. Kevin Zhou. Bridged Semantic Alignment for Zero-shot 3D Medical Image Diagnosis. [CoRR](#), abs/2501.03565, 2025.
- [22] Alon Lavie und Abhaya Agarwal. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In [Proceedings of the Second Workshop on Statistical Machine Translation](#), WMT@ACL 2007, S. 228–231, 2007.
- [23] Cheng-Yi Li, Kao-Jung Chang, Cheng-Fu Yang, Hsin-Yu Wu, Wenting Chen, Hritik Bansal, Ling Chen, Yi-Ping Yang, Yu-Chun Chen, Shih-Pin Chen, Jiing-Feng Lirng, Kai-Wei Chang, und Shih-Hwa Chiou. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. [Nature Communications](#), 16, 2024.

- [24] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, und Jianfeng Gao. LLava-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *Annual Conference on Neural Information Processing Systems 2023*, NeurIPS 2023, 2023.
- [25] Junnan Li, Dongxu Li, Silvio Savarese, und Steven C. H. Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICML 2023*, volume 202, S. 19730–19742, 2023.
- [26] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, und Xiaojun Chang. Dynamic Graph Enhanced Contrastive Learning for Chest X-ray Report Generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, S. 3334–3343, 2023.
- [27] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jianke Zhu, und Lei Zhang. TokenPacker: Efficient Visual Projector for Multimodal LLM. *CoRR*, abs/2407.02392, 2024.
- [28] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, und Jiaya Jia. Mini-Gemini: Mining the Potential of Multi-modality Vision Language Models. *CoRR*, abs/2403.18814, 2024.
- [29] Martens W Hu Y Adriaensens P Quirynen M Lambrichts I Liang X, Jacobs R. Macro-and micro-anatomical, histological and computed tomography scan characterization of the nasopalatine canal. *Journal of clinical periodontology*, 36(7) S. 598–603, 2009.
- [30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, S. 74–81, 2004.
- [31] Xuefeng Ni, Linshan Wu, Jiaxin Zhuang, Qiong Wang, Mingxiang Wu, Varut Vardhanabutti, Lihai Zhang, Hanyu Gao, und Hao Chen. MG-3D: Multi-Grained Knowledge-Enhanced 3D Medical Vision-Language Pre-training. *CoRR*, abs/2412.05876, 2024.
- [32] Kishore Papineni, Salim Roukos, Todd Ward, und Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL 2002*, S. 311–318, 2002.
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, und Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, S. 8748–8763, 2021.
- [34] L. Salvolini, E. B. Secchi, L. Costarelli, et al. Clinical applications of 2D and 3D CT imaging of the airways—a review. *European journal of radiology*, 34(1) S. 9–25, 2000.
- [35] Raphael Sexauer und Caroline Bestler. Time Is Money: Considerations for Measuring the Radiological Reading Time. *Journal of Imaging*, 8, 2022.
- [36] Yiming Shi, Xun Zhu, Ying Hu, Chenyi Guo, Miao Li, und Ji Wu. Med-2E3: A 2D-Enhanced 3D Medical Multimodal Large Language Model. *CoRR*, abs/2411.12783, 2024.
- [37] Ryutaro Tanno, David G. T. Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, Karan Singhal, Mike Schaeckermann, Rhys May, Roy Lee, SiWai Man, S. Sara Mahdavi, Zahra Ahmed, Yossi Matias, Joelle Barral, S. M. Ali Eslami, Danielle Belgrave, Yun Liu, Sreenivasa Raju Kalidindi, Shravya Shetty, Vivek Natarajan, Pushmeet Kohli, Po-Sen Huang, Alan Karthikesalingam, und Ira Ktena. Collaboration between clinicians and vision-language models in radiology report generation. *Nature Medicine*, 31 S. 599 – 608, 2024.
- [38] Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, und Alexey Dosovitskiy. MLP-Mixer: An all-MLP Architecture for Vision. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, NeurIPS 2021, S. 24261–24272, 2021.

- [39] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David J. Fleet, Philip Andrew Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, und Vivek Natarajan. Towards Generalist Biomedical AI. *CoRR*, abs/2307.14334, 2023.
- [40] Aäron van den Oord, Yazhe Li, und Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [41] Jun Wang, Abhir Bhalerao, und Yulan He. Cross-modal prototype driven network for radiology report generation. In *Computer Vision - ECCV 2022 - 17th European Conference*, S. 563–579. Springer, 2022.
- [42] Biao Wu, Yutong Xie, Zeyu Zhang, Minh Hieu Phan, Qi Chen, Ling Chen, und Qi Wu. XLIP: Cross-modal Attention Masked Modelling for Medical Language-Image Pre-Training. *CoRR*, abs/2407.19546, 2024.
- [43] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, und Weidi Xie. Towards Generalist Foundation Model for Radiology. *CoRR*, abs/2308.02463, 2023.
- [44] Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, Wenhao Zheng, Zhaoyang Wang, Xiao Wang, Xuchao Zhang, Chetan Bansal, Marc Niethammer, Junzhou Huang, Hongtu Zhu, Yun Li, Jimeng Sun, Zongyuan Ge, Gang Li, James Y. Zou, und Huaxiu Yao. CARES: A Comprehensive Benchmark of Trustworthiness in Medical Vision Language Models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*, 2024.
- [45] Yu Xin, Gorkem Can Ates, Kuang Gong, und Wei Shao. Med3DVLM: An Efficient Vision-Language Model for 3D Medical Image Analysis. volume *arXiv:2503.20047*, 2025.
- [46] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, und Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, S. 11941–11952, 2023.
- [47] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, und Hoifung Poon. A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. *NEJM AI*, 2(1), 2024.
- [48] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, und Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.
- [49] Xiaodan Zhang, Yanzhao Shi, Junzhong Ji, Chengxin Zheng, und Liangqiong Qu. MEPNet: Medical Entity-Balanced Prompting Network for Brain CT Report Generation. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*, S. 25940–25948, 2025.
- [50] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, und Weidi Xie. RadGenome-Chest CT: A Grounded Vision-Language Dataset for Chest CT Analysis. *CoRR*, abs/2404.16754, 2024.
- [51] Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, und Weidi Xie. RaTEScore: A Metric for Radiology Report Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, S. 15004–15019, 2024.

## A Visual Token Compression Sensitivity Study

To further explore the token compression and 3D spatial preservation capabilities of our spatial packer, we perform sensitivity experiments comparing various strides ( $S_d, S_w, S_h$ ) to control the number of compressed tokens during the encoding of low-resolution points  $\mathbf{V}_{lr}^G \in \mathbb{R}^{S_d \times S_w \times S_h \times d_v}$  (See section 3.3 in the manuscript).  $S_d, S_w$ , and  $S_h$  denote the count of partitioned voxels in the spatial dimensions of the volume feature,  $\hat{D}, \hat{W}$ , and  $\hat{H}$ , with each voxel having dimensions  $(\frac{\hat{D}}{S_d}, \frac{\hat{W}}{S_w}, \frac{\hat{H}}{S_h})$ .

Table 5: Performance comparison across different numbers of visual tokens in the spatial packer for report generation. Ratios (e.g., X% $\downarrow$  or Y% $\uparrow$ ) indicate the degree of token reduction or expansion relative to the default setting.

Token Number	Stride	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-L	METEOR	BERT-Score	RaTE-Score
32 (75% $\downarrow$ )	(8,2,2)	60.95	36.37	25.33	19.77	55.96	37.78	43.92	88.69	63.16
64 (50% $\downarrow$ )	(4,4,4)	61.43	36.73	25.66	20.11	55.92	37.96	43.72	88.73	63.40
128 (default)	(8,4,4)	62.89	<b>39.47</b>	<b>29.11</b>	<b>24.01</b>	<b>56.50</b>	<b>40.63</b>	<b>44.75</b>	<b>88.99</b>	<b>64.99</b>
256 (100% $\uparrow$ )	(4,8,8)	<b>63.41</b>	38.31	27.09	21.59	55.73	38.80	43.06	88.84	63.55

As shown in Table 5, when the number of visual tokens is compressed to 32, the model performance decreases compared to our original configuration (128 tokens per spatial packer), with BLEU-1 and ROUGE-L dropping by 1.94% and 2.85%, respectively. Nonetheless, its clinical performance is still equivalent to the variant of spatial pooling-based projector with 128 tokens (RaTE-Score: 63.16%, see Table 2(b) in our main manuscript), which demonstrates that our spatial packer can preserve the clinical relevance of generated context, even under extreme token compression. As the visual token count increases from 32 to 128 (32  $\rightarrow$  64  $\rightarrow$  128), we observe a gradual improvement in performance, particularly in BLEU-4 (from 19.77% to 20.11% to 24.01%, 4.24%  $\uparrow$  in total). This suggests that the model’s ability to generate coherent and contextually accurate text improves with more visual tokens, emphasizing the importance of token quantity for text generation quality. However, increasing the token count beyond 128, particularly to 256, results in degraded performance (1.44% $\downarrow$  of RaTE-Score). The reason may be due to the reduced voxel size  $(\frac{\hat{D}}{S_d}, \frac{\hat{W}}{S_w}, \frac{\hat{H}}{S_h})$ , which impairs the V2P-CA module’s capacity to capture salient high-resolution structures. Consequently, the spatial features become less discriminative, leading to decreased overall performance.

## B Evaluations on BIMCV-R Dataset

We conduct additional experiments on the BIMCV-R dataset [10], a benchmark for 3D medical report generation. This dataset comprises 8,069 3D CT volumes (over 2 million slices), each paired with a corresponding medical report. Following the preprocessing protocol of Lai et al. [20], we use 6,766 volume-report pairs for training and 752 for testing. We reuse our dual 3D visual encoders pretrained on the CT-RATE dataset [14] without further updates to extract volume features from BIMCV-R. Only the spatial packer and LoRA layers are fine-tuned for task adaptation.

The results of report generation are presented in Table 6. Notably, despite using frozen visual encoders ( $\mathbf{E}_{3d}(\cdot)$  and  $\mathbf{E}_{2e3}(\cdot)$ ) pretrained exclusively on the CT-RATE dataset, HSENet achieves the best performance across all evaluation metrics, including a 14.28% increase in BLEU-1 over E3D-GPT. This demonstrates the effectiveness of our pretraining strategy in capturing valuable spatial patterns from 3D CT volumes. It is also interesting to find that E3D-GPT, which adopts self-reconstruction for visual pretraining, obtains the second-best results in BERTScore, ROUGE-1, and METEOR (81.78%, 23.93%, and 13.62%, respectively). This suggests that, under limited data conditions (BIMCV-R contains only 14.3% as many training samples as CT-RATE), self-reconstruction may enable the learning of more expressive medical representations than CLIP-style pretraining, thus benefiting downstream report generation. These findings point to a promising direction for future research: integrating self-reconstruction with vision-language alignment to further enhance the understanding of 3D medical visual features.

## C Evaluation of Clinical Efficiency in VQA

We evaluate the clinical efficiency of HSENet in answering questions across various anatomical locations, including the heart, breast, and lung. Ten key body locations are selected based on the

Table 6: Experiments on medical report generation on the BIMCV-R dataset [10]. **Bold** indicates the best performance.  $\dagger$  denotes the reproduced models.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-L	METEOR	BERT-Score	RaTE-Score
RadFM[43]	0.83	/	/	/	3.87	/	1.98	78.21	/
CT-CHAT[14]	/	/	/	/	/	/	/	/	/
M3D-LaMed[3]	16.43	/	/	/	21.44	/	11.38	81.63	/
E3D-GPT[20]	18.19	/	/	/	23.93	/	13.62	81.78	/
Med-2E3[36] $\dagger$	27.32	5.99	2.01	0.79	14.77	10.55	8.40	80.01	34.65
Med3DVLM[45] $\dagger$	31.13	5.29	1.55	0.66	20.09	12.05	11.55	81.73	32.92
HSENet (Ours)	<b>32.47</b>	<b>7.33</b>	<b>2.71</b>	<b>1.43</b>	<b>24.88</b>	<b>14.98</b>	<b>14.67</b>	<b>82.50</b>	<b>36.13</b>

official categories provided by RadGenome-ChestCT dataset [50]. We compare HSENet against several strong baselines, including M3D-LaMed [3], Med-2E3 [36], and Med3DVLM [45], as well as different variants of HSENet using single or dual visual encoders.

As shown in the final subfigure of Figure 5, HSENet achieves the highest overall F1 score (79.42%) among all methods, demonstrating its ability to interpret diverse 3D spatial patterns in complex clinical reasoning tasks. It performs especially promising on anatomically stable regions such as the heart (90.48%), lung (94.88%), and breast (65.47%). In contrast, we also noticed that the performance on structurally irregular regions like bone is less optimal. Despite a strong F1 score of 94.82%, HSENet slightly underperforms Med-2E3 by 0.35%. This may be attributed to the uniform voxel partitioning used in our spatial packer, which limits its adaptability to highly variable skeletal structures. Therefore, introducing adaptive voxel partitioning could offer a promising future direction for enhancing spatial encoding and improving performance in regions with complex anatomical variation.

## D Training, Inference, and Computational Resources

This section outlines the detailed configurations and computational resources used for model training and inference.

**Vision-Language Pretraining.** We perform two-stage pretraining on the CT-RATE dataset [14]. Both pretraining stages are trained for 50 epochs using 8 NVIDIA RTX 3090 GPUs, with approximately 23 GB of memory use and 24 data loader workers per GPU. Each stage requires roughly 26-28 hours of training.

**MLM Fine-tuning.** We apply 8-bit quantization to the LLM and fine-tune it with LoRA [15].

Fine-tuning is conducted on 8 NVIDIA RTX 3090 GPUs. For the report generation task, we train on the CT-RATE dataset [14] for 6 epochs, using 22 GB memory per GPU and 22 workers per GPU, requiring approximately 14 hours. For the VQA task, we train on the RadGenome-ChestCT dataset [50] for 4 epochs with similar resource settings, taking about 22 hours in total.

**Inference Efficiency.** Table 7 presents the inference latency of HSENet. Our HSENet generates diagnostic reports in 3.59 seconds per instance and answers VQA queries in 1.11 seconds on average. Compared to human radiologists, who require approximately 950.40 seconds (15.84 minutes) per report [35], HSENet achieves a  $\sim 264\times$  speedup in report generation.

## E Additional Qualitative Analysis

**3D Patch Scoring.** Figure 6 shows the scoring distributions of 3D patches generated by our 2E3 Visual Encoder  $E_{2e3}(\cdot)$ . These distributions exhibit substantial variation across samples, suggesting that our model effectively captures both intra-sample patch differences and inter-sample visual variability. This adaptive scoring strategy can effectively enhance the discriminability of learned 3D visual representations, thereby boosting the model’s performance in multi-modal retrieval and text generation tasks.

Table 7: Results of inference efficiency analysis. *s/item* refers to seconds per item. The inference time for human radiologists is provided from Sexauer und Bestler [35] for reference.

	Report Generation	VQA
Ours	3.59 <i>s/item</i>	1.11 <i>s/item</i>
Radiologist	$\sim 950.40$ <i>s/item</i>	/

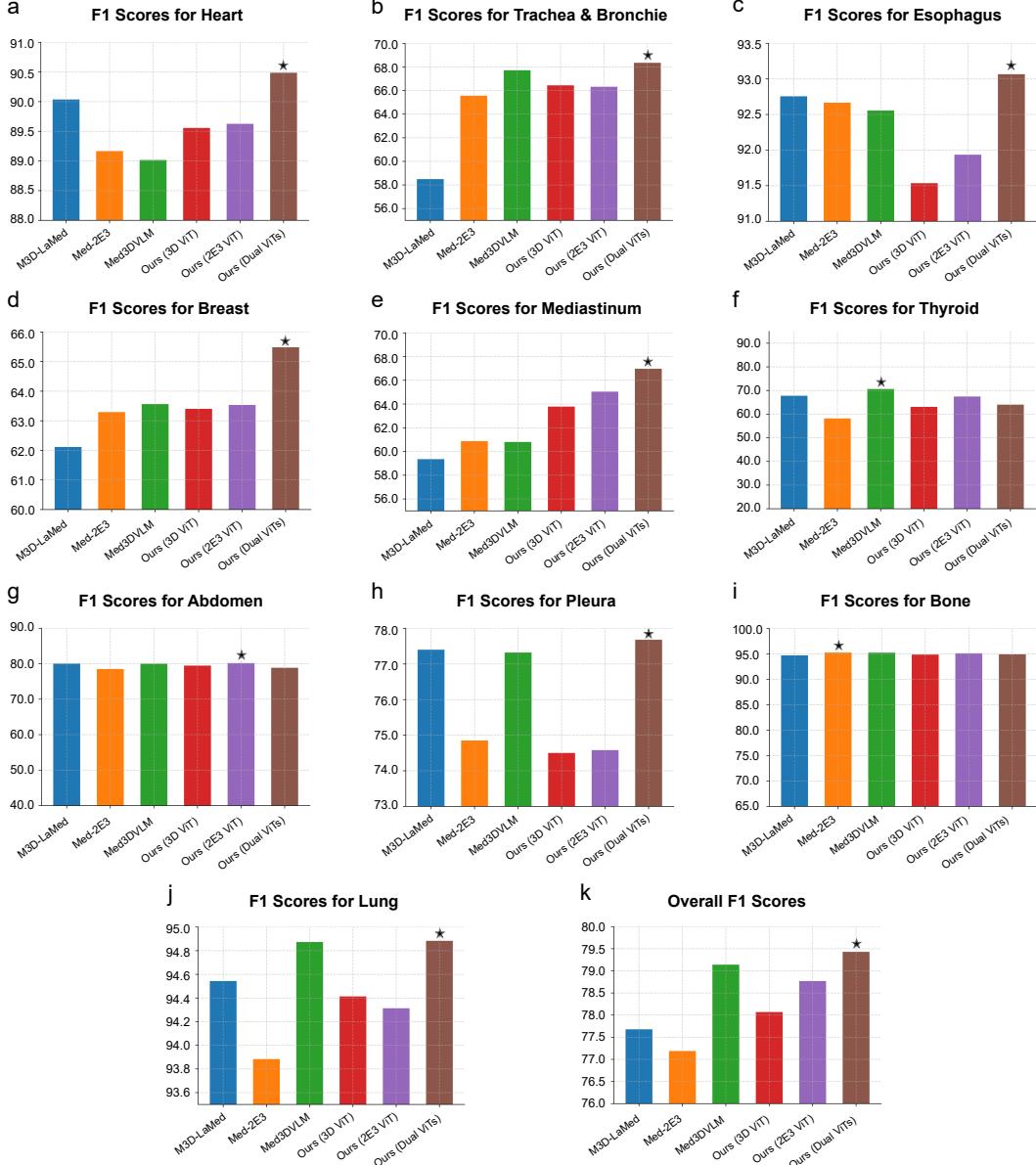


Figure 5: Experiments on the clinical effectiveness of VQA across different body locations. Colored bars represent different methods, with the star symbol (\*) indicating the highest F1 score. *3D ViT*, *2E3 ViT*, and *Dual ViTs* denote the use of our 3D Visual Encoder  $E_{3d}(\cdot)$ , 2E3 Visual Encoder  $E_{2e3}(\cdot)$ , and both encoders within the proposed HSENet, respectively. Each subfigure records the F1 score for a specific body location, while the final subfigure (k) shows the average clinical performance across all locations.

**Volume-Report Retrieval.** To evaluate retrieval performance, we provide qualitative examples using the pretrained stage-2 multi-modal encoders, i.e., 2E3 visual encoder  $E_{2e3}(\cdot)$  and text decoder  $E_{text}^{s_2}(\cdot)$ , to extract features from 3D volumes and medical reports. As shown in Figure 7, given a query volume, our method retrieves its ground-truth report from a test set of 3,039 volume-report pairs with high confidence (0.979). Notably, the top-2 and top-3 retrieved reports also demonstrate strong semantic similarity to the ground truth, despite minor lexical variations (e.g., “There are emphysematous changes in both lungs.” vs. “Emphysematous changes are observed in both lungs.”). These results indicate that our 2D-enhanced 3D learning framework effectively captures cross-modal correlations, enabling accurate and semantically aligned volume-report retrieval.

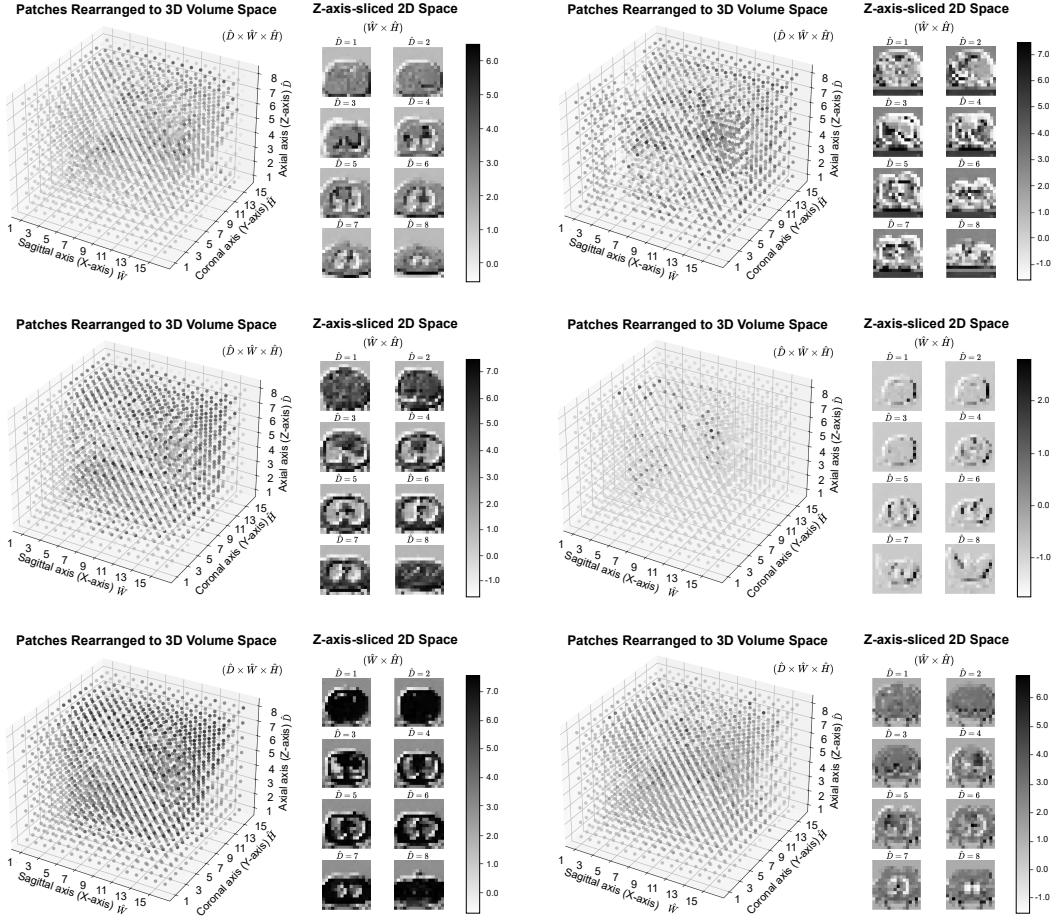


Figure 6: Additional visualizations of 3D patch scores generated by the 2E3 Visual Encoder  $E_{2e3}(\cdot)$ . Darker colors indicate higher scores. Both 3D views (patches rearranged into the original volume space) and 2D views (axial slices along the Z-axis at different depth levels  $\hat{D}$ ) are provided to illustrate the spatial distribution of scores.

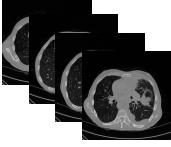
## F Textual Prompts for Model Training

To empower HSENet with instruction-following capabilities towards 3D medical tasks, e.g., medical report generation and medical VQA, we adopt a diverse set of textual prompts during training. For report generation, we follow the protocol of Bai et al. [3], utilizing 42 distinct prompt templates (see Figure 8). A prompt is randomly selected for each training instance to improve the robustness and generalization of the vision-language model. These prompts are consistently applied across ablation studies and baseline comparisons to ensure fairness. For the VQA task, we similarly employ 50 prompt types from Zhang et al. [50] (see Figure 9), enabling HSENet to generalize across a wide range of question formats.

## G Limitations

Clinical diagnosis typically relies on a combination of 3D visual data and rich contextual information, including patient history, clinical interviews, and electronic health records. While this work tackles a core challenge of learning generalizable 3D spatial representations and yields strong performance across a range of downstream tasks, we do not explicitly address the organization or integration of diverse contextual clinical data during pretraining or fine-tuning. This omission may lead to suboptimal diagnostic text generation in more complex, real-world scenarios, potentially undermining

### Medical Volume-Report Retrieval



**Ground Truth:** A mass measuring 3 cm is observed in the thickest part of the right lung, which completely surrounds the pleura at its apex. Between the pleural leaves on the right, there are effusion areas measuring 53 mm in the thickest part and showing loculation in places. In the upper lobe of the right lung, reticular density increases with irregular borders were observed and were evaluated as compatible with lymphangitic spread. In addition, there is a consolidation area in the middle lobe with air bronchograms and atelectatic changes. There are irregular thickenings in the mediastinal and costal pleura. Soft tissue densities are observed in the lower paratracheal area, approximately 36x30 mm in size, with a central necrotic appearance and conglomerate lymphadenopathy. In addition, there are central necrotic lymphadenopathies in the upper-lower paratracheal, subcarinal paraesophageal and right hilar areas, the largest of which measures 3 cm on the short axis. Emphysematous changes are observed in both lungs. There is parenchymal fibrosis and bulla formation in the upper lobe of the left lung causing volume loss. Millimetric parenchymal nodules are observed in the upper and lower lobes of the left lung. A 5 mm diameter parenchymal nodule was observed in the middle lobe of the right lung. In the upper abdominal organs included in the sections, there are lymphadenopathies measuring 27x17 mm in size at the level of the celiac and superior mesenteric arteries. Bone structures in the study area are natural. Vertebral corpus heights are preserved.

**Top 1 (Similarity: 0.979):** A mass measuring 3 cm is observed in the thickest part of the right lung, which completely surrounds the pleura at its apex. Between the pleural leaves on the right, there are effusion areas measuring 53 mm in the thickest part and showing loculation in places. In the upper lobe of the right lung, reticular density increases with irregular borders were observed and were evaluated as compatible with lymphangitic spread. In addition, there is a consolidation area in the middle lobe with air bronchograms and atelectatic changes. There are irregular thickenings in the mediastinal and costal pleura. Soft tissue densities are observed in the lower paratracheal area, approximately 36x30 mm in size, with a central necrotic appearance and conglomerate lymphadenopathy. In addition, there are central necrotic lymphadenopathies in the upper-lower paratracheal, subcarinal paraesophageal and right hilar areas, the largest of which measures 3 cm on the short axis. Emphysematous changes are observed in both lungs. There is parenchymal fibrosis and bulla formation in the upper lobe of the left lung causing volume loss. Millimetric parenchymal nodules are observed in the upper and lower lobes of the left lung. A 5 mm diameter parenchymal nodule was observed in the middle lobe of the right lung. In the upper abdominal organs included in the sections, there are lymphadenopathies measuring 27x17 mm in size at the level of the celiac and superior mesenteric arteries. Bone structures in the study area are natural. Vertebral corpus heights are preserved.

**Top 2 (Similarity: 0.935):** Heart contour and size are normal. Pericardial effusion was not detected. There are stent formations in the anterior descending coronary artery. Calcific atheroma plaques are observed in the aorta. The widths of the mediastinal main vascular structures are normal. Multiple FDG positive lymph nodes with 11 mm diameter are observed in the mediastinum and bilateral hilar regions, the largest in the prevascular area. Trachea and both main bronchi are open. No occlusive pathology was detected in the trachea and both main bronchi. In a patient who underwent pleurectomy and diaphragmatic resection due to mesothelioma, a primary mass characterized by plaque-like nodular pleural thickness increase whose borders cannot be distinguished from the mediastinum in the medial direction from the upper lobe of the right lung to the lower lobe, and postoperative hyperdense material on the diaphragm face are observed. It is observed that the mass extends under the skin from the intercostal area in the anterior part of the 6th rib. In the upper lobe of the right lung, there is a consolidation area in which air bronchograms are observed and sometimes accompanied by ground glass. In the middle lobe and lower lobe of the right lung, diffuse parenchymal soft tissue lesions and accompanying ground-glass areas are observed. Multiple metastatic nodules of 10x12 mm are observed in both lungs, the largest of which is in the superior segment of the left lung lower lobe. There are occasional millimetric parenchymal air cysts in the left lung. There are areas of linear atelectasis in the left lung apicoposterior segment and lower lobe posterior segment. Sliding type hiatal hernia is observed at the esophagogastric junction. As far as it can be evaluated within the limits of non-contrast CT; There are millimetric nodular metastatic lesions in the capsular area at the level of the posterior segment of the right lobe of the liver. A view compatible with the omental cake is observed. No lytic-destructive lesions were observed in the bone structures within the sections. In the lateral-posterior wall of the right thorax, there are multiple nodular metastatic lesions, the largest measuring 16x20 mm, within the subcutaneous fatty tissue and muscle planes.

**Top 3 (Similarity: 0.875):** In the left hemithorax, at the level of the 2nd-5th ribs, an appearance of soft tissue density is observed, with a clear borderless infiltrative character extending from the intercostal spaces to the outside of the hemithorax. The described view measures 32 mm at its thickest point (series 2 section 203). This appearance was evaluated primarily in favor of the mass. No significant destruction was detected in the ribs. There is pleural effusion on the left. The pleural effusion measured 53 mm at the level of the lower lobe of the lung at its thickest point. The described view measured approximately 20 mm at its thickest point. The described appearance could not be characterized because no contrast medium was given. However, when evaluated together with other findings, there may be a soft tissue mass in this appearance. Further investigation is recommended. No pleural effusion or thickening was detected on the right. There are lymphadenopathies in the left axilla and retropectoral region. The shortest diameter of the largest lymphadenopathy described was 19 mm at its widest point (series 2 section 76). No pathologically enlarged lymph nodes were detected in the right axilla and retropectoral region. There are millimetric lymph nodes in the left internal mammary artery trace. Lymphadenopathy with a short diameter of 26mm was observed in the subcarinal area. In addition, there are millimetric lymph nodes in the mediastinum and hilar regions. There is no obstructive pathology in the trachea and both main bronchi. In the central part of the lower lobe of the left lung, there is consolidation with an air bronchogram. This appearance was primarily evaluated in favor of infective pathology. However, when evaluated together with other findings, this appearance may also belong to a metastatic mass. This distinction cannot be made in this examination. It is recommended to be evaluated together with previous examinations, if any. Ground glass areas are also present in the lower lobe of both lungs and the upper lobe of the left lung. Ground glass areas are more prominent in the lower lobes. These views are nonspecific. There are emphysematous changes in both lungs. No mass or infiltrative lesion was detected in the right lung. There are millimetric nodules in both lungs. The appearance of the described nodules is also non-specific. The largest of the nodules is observed in the lower lobe of the right lung and its longest diameter is approximately 9 mm. No upper abdominal free fluid-collection was detected in the sections. No pathologically enlarged lymph nodes were observed. There are no lytic-destructive lesions in the bone structures within the sections.

Figure 7: Visualization of medical volume-to-report retrieval. The 2E3 visual encoder  $E_{2e3}(\cdot)$  and the text decoder  $E_{text}^{s2}(\cdot)$  is utilized to encode 3D volume and report features, respectively. For each input volume, the top-3 retrieved reports are shown to assess retrieval quality. Underlined sentences highlight key findings consistent with the ground-truth report.

clinical reliability. Therefore, a key direction for future work is the effective collection and integration of multi-modal, multi-source clinical data to improve the robustness and reliability of 3D diagnostic systems.

## H Dataset License

This work uses publicly available benchmark datasets: CT-RATE [14] (CC-BY-NC-SA 4.0 License), RadGenome-ChestCT [50] (CC-BY 4.0 License), and BIMCV-R [10] (MIT License). All licenses permit usage for research purposes. We fully comply with the respective license terms, and all datasets are used solely for research without any modification or repackaging.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction state the scope and contributions of the paper, including the introduction of the HSENet, its effective pretraining strategy and spatial packer for visual perception and projection, and the substantial performance gains achieved across diverse downstream tasks.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper includes a section on limitations in supplemental material, outlining the scope of the framework, and areas for future improvement.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not present theoretical results that necessitate formal assumptions or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides comprehensive details on datasets, experimental setups, and methodologies used, ensuring that the results can be reproduced accurately.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is included in the supplemental material and will be open-sourced upon acceptance to support the 3D medical vision-language understanding research community.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies relevant experimental details, including data splits, number of samples, and hyperparameters, ensuring transparency and reproducibility of the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars are not reported due to the high computational cost of 3D medical volume-report modeling.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Detailed information on computational resources, including workers, memory, and inference time, is provided in the supplementary materials to ensure reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: This research aligns with the NeurIPS Code of Ethics, ensuring responsible conduct throughout the study.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses potential societal impacts, with positive impacts mentioned in the introduction, and negative impacts mentioned in the limitations section (see supplement materials). These include the benefits of medical-assisted models as well as risks related to medical hallucinations in generated responses.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The paper discusses potential limitations in supplemented materials, such as risks of medical hallucinations and incorrect diagnoses in 3D medical image analysis.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper properly credits the creators of existing assets used and states the licenses and terms of use.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](http://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not introduce any new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The LLM is utilized in the HSENet as the language decoder, which is not the core innovation of this research. This is clarified in the method and experiment sections.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

## Medical Report Generation Prompts

- Can you provide a caption consists of findings for this medical image?
- Describe the findings of the medical image you see.
- Please caption this medical scan with findings.
- What is the findings of this image?
- Describe this medical scan with findings.
- Please write a caption consists of findings for this image.
- Can you summarize with findings the images presented?
- Please caption this scan with findings.
- Please provide a caption consists of findings for this medical image.
- Can you provide a summary consists of findings of this radiograph?
- What are the findings presented in this medical scan?
- Please write a caption consists of findings for this scan.
- Can you provide a description consists of findings of this medical scan?
- Please caption this medical scan with findings.
- Can you provide a caption consists of findings for this medical scan?
- Please generate a medical report based on this image.
- Can you generate a diagnose report from this image.
- Could you analyze and provide a caption for the findings in this medical image?
- Please describe the observations depicted in this medical scan.
- Can you summarize the findings of this image in a caption?
- What are the significant findings in this medical image?
- Please provide a detailed caption outlining the findings of this image.
- Could you interpret and describe the findings shown in this medical scan?
- What conclusions can you draw from the observations in this image?
- Please write a descriptive caption based on the findings in this scan.
- What key findings can you identify from examining this medical image?
- Could you generate a detailed report based on the observations in this image?
- Can you provide a diagnosis based on the findings in this image?
- Please generate a comprehensive report summarizing the findings in this image.
- Caption the findings in this medical image?
- Describe the findings you see.
- Caption this medical scan's findings.
- What are the findings here?
- Describe these findings.
- Summarize the findings in these images.
- Caption this scan's findings.
- Provide a caption for this medical image's findings.
- Summarize the findings of this radiograph.
- What findings are presented in this scan?
- Describe this scan's findings.
- Generate a medical report based on this image.
- Can you provide a diagnosis based on this image?

Figure 8: Textual prompts for medical report generation follow the format of Bai et al. [3]. To enhance the instruction-following capability of HSENet, prompts are randomly assigned to samples during training.

## Medical VQA Prompts

- Where is the *{abnormality}* located in the image?
- Where can the *{abnormality}* be found within the image?
- Where in the image is the *{abnormality}* located?
- Where in the image is the *{abnormality}* localized?
- Where in the image can the *{abnormality}* be found?
- Where in the image does the *{abnormality}* appear?
- Where in the image does the *{abnormality}* locate?
- Where in the image does the *{abnormality}* locate?
- Where specifically within the image is the *{abnormality}* located?
- Where exactly within the image is the *{abnormality}* located?
- Where exactly is the *{abnormality}* located in the image?
- Where specifically is the *{abnormality}* located in the image?
- Where exactly within the image is the *{abnormality}* localized?
- Where specifically within the image is the *{abnormality}* localized?
- Where within the image can the *{abnormality}* be precisely located?
- Where exactly within the image does the *{abnormality}* present?
- Where within the image does the *{abnormality}* specifically present?
- Where in the image does the *{abnormality}* appear?
- What is the location of the *{abnormality}* in the image?
- What is the precise location of the *{abnormality}* in the image?
- What is the specific location of the *{abnormality}* within the image?
- What is the precise region of the *{abnormality}* in the image?
- What is the specific region of the *{abnormality}* within the image?
- What particular region within the image does the *{abnormality}* occupy?
- What particular location within the image does the *{abnormality}* occupy?
- What specific location within the image does the *{abnormality}* occupy?
- What specific region within the image does the *{abnormality}* occupy?
- What specific area of the image does the *{abnormality}* occupy?
- What specific region of the image does the *{abnormality}* appear?
- What specific spot within the image contains the *{abnormality}*?
- What particular region of the image is affected by the *{abnormality}*?
- What specific area within the image is impacted by the *{abnormality}*?
- What specific region within the image is impacted by the *{abnormality}*?
- What specific location within the image is impacted by the *{abnormality}*?
- What particular region within the image is affected by the *{abnormality}*?
- What particular area within the image is affected by the *{abnormality}*?
- What particular location within the image is affected by the *{abnormality}*?
- What specific region within the image does the *{abnormality}* affect?
- What specific area within the image does the *{abnormality}* affect?
- What specific location within the image does the *{abnormality}* affect?
- What specific location within the image does the *{abnormality}* appear?
- What specific region within the image does the *{abnormality}* appear?
- What specific area within the image does the *{abnormality}* appear?
- What particular spot within the image does the *{abnormality}* present?
- What particular area within the image does the *{abnormality}* present?
- What particular region within the image does the *{abnormality}* present?
- What specific area within the image does the *{abnormality}* occur?
- What specific location within the image does the *{abnormality}* occur?
- What specific region within the image does the *{abnormality}* occur?

Figure 9: Textual prompts for medical VQA follow the format of Zhang et al. [50]. To ensure HSENet’s instruction-following ability, prompts are randomly assigned to training samples. The placeholder *{abnormality}* indicates where location-specific abnormalities are inserted.