# MedVTok: A General-Purpose Medical Visual Tokenizer

**Chenglong Ma[1,2,*], Yuanfeng Ji[3,7,*], Junzhi Ning[4,*], Jiyao Liu[1,*], Yujie Wei[1], Shaohao Rui[2,6], Chaoyang Zhang[2], Wenjie Li[2,6], Wenhao Tang[4], Cheng Tang[4], Wei Li[4,6], Shujian Gao[1,4], Yujie Zhang[1,2], Ying Chen[4], Zilong Li[1], Zongyuan Ge[5], Diping Song[4], Jin Ye[4,5], Ming Hu[4,5], Junjun He[2,4,7,†], Hongming Shan[1,†]**

[1]Fudan University, China [2]Shanghai Innovation Institute, China
[3]Stanford University School of Medicine, United States
[4]Shanghai Artificial Intelligence Laboratory, China [5]Monash University, Australia
[6]Shanghai Jiao Tong University, China [7]Ant Healthcare, Ant Group, China

clma24@m.fudan.edu.cn yfj@stanford.edu

## Abstract

Multimodal medical AI requires a common visual interface for encoding heterogeneous imaging evidence for patient condition modeling. However, existing medical visual tokenizers are often tied to a specific dimension and imaging modality, forcing multimodal systems to rely on *fragmented and poorly reusable* representations. We introduce MedVTok, a general-purpose visual tokenizer that maps 2D images and 3D volumes from diverse imaging modalities into a unified token space. Building such a tokenizer poses two key challenges: preserving anatomical structure across dimensions and capturing clinical semantics across modalities. To this end, MedVTok introduces (**i**) slice differential regularization, which explicitly models inter-slice anatomical coherence missing in previous per-pixel/per-voxel supervision, and (**ii**) multi-expert representation alignment, which integrates knowledge from multiple medical expert encoders while retaining modality-specific complementary knowledge. Trained on 28M images and 70K volumes across 9 imaging modalities, MedVTok supports classification, retrieval, segmentation, synthesis, visual question answering, and medical report generation through a *single* tokenizer. Across 24 benchmarks, MedVTok achieves state-of-the-art performance, demonstrating a scalable token interface that unifies medical image dimension, modality, and task usage. Project page: https://masaaki-75.github.io/medvtok

## 1 Introduction

Multimodal medical AI aims to integrate heterogeneous patient evidence, including laboratory measurements, longitudinal records, and medical images, to form a holistic understanding of a patient's condition. As the field shifts from isolated task-specific models toward generalist multimodal systems [1–3], a central requirement is *a shared interface through which diverse clinical evidence can be encoded, transferred, and reused across tasks*. Among these evidence sources, medical images are particularly rich and high-dimensional, capturing anatomical structures, pathological findings, and disease progression. Visual tokenization therefore serves as a critical front end that converts raw imaging data into compact visual tokens compatible with multimodal models.

*These authors contributed equally to this work.

†Co-corresponding authors. Hongming Shan is affiliated with the Institute of Science and Technology for Brain-inspired Intelligence (ISTBI), the MOE Frontiers Center for Brain Science, and the Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Ministry of Education), Fudan University, China.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

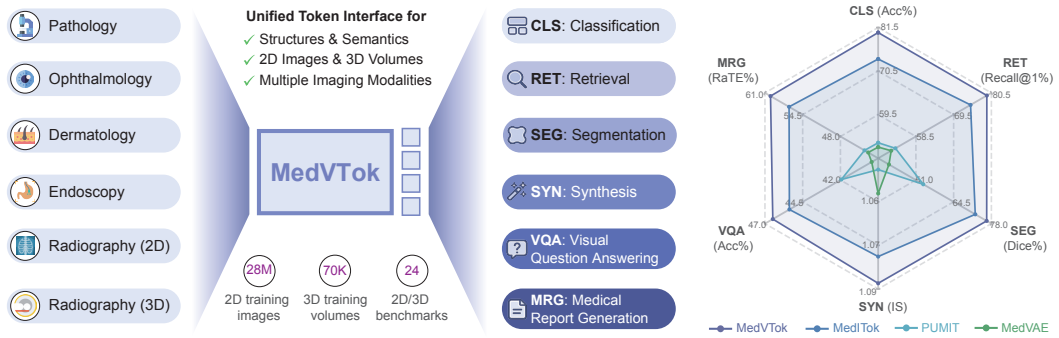



Figure 1: **MedVTok provides a unified token interface for multimodal modeling with medical imaging data.** Trained on 28M images and 70K volumes, MedVTok achieves state-of-the-art performance across a broad range of downstream tasks covering 24 2D/3D benchmarks.

However, medical imaging data are highly heterogeneous across both modality and dimension, including 2D images such as dermoscopic images, color fundus photographs (CFP) and histopathology patches, as well as 3D volumes such as computed tomography (CT) and magnetic resonance imaging (MRI). These data differ in acquisition protocol, anatomical coverage, and diagnostic content. Existing medical visual tokenizers are often tied to a specific modality and dimension, producing fragmented representations that are difficult to reuse in a multimodal system. Consequently, downstream models must absorb much of the burden of modality adaptation and cross-modality representation learning [4]. A general-purpose medical visual tokenizer should instead provide a single token interface across modalities while natively supporting both 2D images and 3D volumes.

Building such a tokenizer poses two key challenges. *(i) Preserving structural fidelity across dimensions.* Unlike 2D images, 3D scans present volumetric continuity, where organs, vessels, lesions, and tissue boundaries extend across adjacent slices. Existing tokenizers often specialize in either 2D or 3D settings [5, 6], or are optimized for local appearance fidelity or per-pixel/per-voxel reconstruction [7, 8], failing to encode inter-slice structure. *(ii) Capturing multimodal semantics across modalities.* Different modalities may describe related clinical concepts while exposing distinct evidence, *e.g.*, histopathology reveals cellular and tissue-level morphology, whereas CT and MRI capture organ-level anatomy. Existing semantic alignment strategies typically learn representations of different imaging modalities from a single pretrained semantic encoder [6, 9–11], which may underrepresent modality-specific semantics outside that encoder's representation space.

In this paper, we introduce MedVTok, a general-purpose medical visual tokenizer that maps heterogeneous 2D images and 3D volumes from diverse imaging modalities into a unified token space. To preserve structural fidelity across dimensions, MedVTok introduces *Slice Differential Regularization* (SliDR) in addition to per-pixel/per-voxel supervision, which explicitly models inter-slice volumetric continuity. To capture clinical semantics across modalities, MedVTok proposes *Multi-Expert Representation Alignment* (MERA), which builds a shared representation space integrating knowledge from multiple modality-specialized experts instead of a single one, while retaining complementary semantics in each expert's representation space. Together, they enable MedVTok to produce visual tokens that preserve structural fidelity and encode multimodal clinical semantics.

To train and evaluate MedVTok at scale, we curate a large-scale medical imaging corpus spanning 9 modalities, with over 28M 2D images and 70K 3D volumes. We assess MedVTok across 6 task types: classification, segmentation, retrieval, synthesis, visual question answering, and medical report generation, covering broad utility across 24 benchmarks. Our results show that MedVTok achieves state-of-the-art performance, demonstrating the value of a scalable token interface that unifies medical image dimension, modality, and downstream task usage.

**Contributions.** Our main contributions are as follows. ***(i)*** We introduce MedVTok, a general-purpose medical visual tokenizer that provides a shared token interface across diverse imaging modalities and natively supports both 2D images and 3D volumes. ***(ii)*** We propose a multi-level supervision strategy that combines slice differential regularization for inter-slice continuity with multi-expert representation alignment for cross-modality clinical semantics. ***(iii)*** We curate a large-scale medical imaging corpus and conduct comprehensive evaluations across 24 benchmarks, demonstrating the

broad utility of MedVTok across classification, segmentation, retrieval, synthesis, visual question answering, and medical report generation.

## 2 Related Work

**Tokenization across 2D and 3D medical imaging.** Visual tokenizers convert images into compact tokens while preserving structural fidelity [7, 12]. Recent tokenizers [9, 13] further explore how a single tokenizer can handle heterogeneous visual formats such as images and videos. In medical imaging, however, tokenizers are commonly tied to specific input dimensionalities [5, 6]. A few works have explored adaptation to different dimensionalities by pooling 3D weights into 2D [8] or by duplicating 2D images to fit a 3D model [14]. While these approaches can handle 2D and 3D inputs simultaneously, they do not consider the respective properties of each dimensionality, *e.g.*, through-plane anatomical continuity in volumes, and thus cannot fully preserve structural fidelity across input forms. In contrast, MedVTok introduces Slice Differential Regularization to explicitly preserve anatomical continuity in volumetric data, and provides a unified 2D–3D token interface via dimension-agnostic patchification.

**Semantic representation across medical imaging modalities.** Medical image representation learning has produced both broad biomedical encoders and modality-specialized expert models. General medical encoders [15–19] learn transferable representations from large-scale biomedical data, offering broad coverage across medical domains but potentially missing specialty-specific cues. Meanwhile, specialized encoders capture finer clinical semantics within individual imaging modalities or specialties [14, 20–26], yet remain isolated in different representation spaces. Existing tokenizers [5, 6] have only partially captured these different semantics, *e.g.*, MedITok [6] aligns its representations with those of a single semantic encoder [15], thereby inheriting the single-expert bias and limiting the learning of complementary knowledge across modalities. MedVTok instead uses Multi-Expert Representation Alignment to integrate knowledge from multiple medical experts into a single token space, while retaining modality-specific semantics.

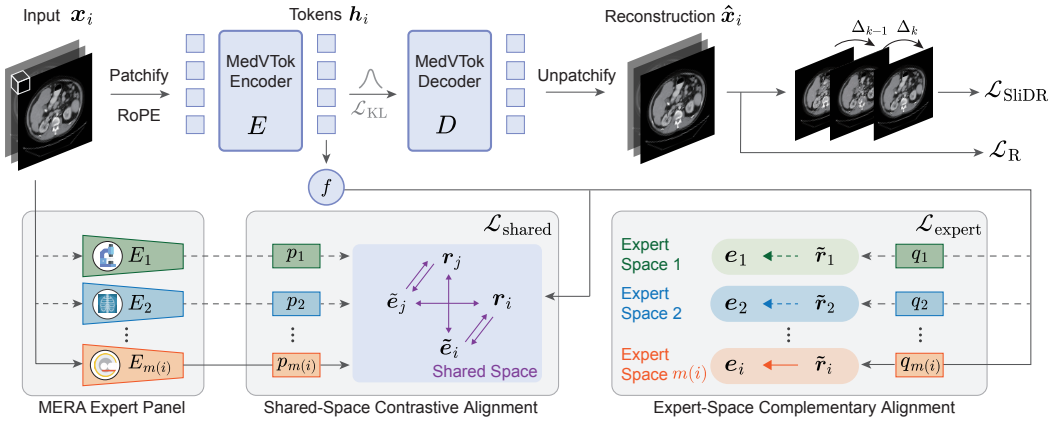



Figure 2: **Overview of the MedVTok training framework.** MedVTok learns both (1) structural fidelity across 2D and 3D inputs under the reconstruction loss $\mathcal{L}_{\mathrm{R}}$ plus the slice differential regularization $\mathcal{L}_{\mathrm{SliDR}}$, and (2) clinical semantics across modalities driven by $\mathcal{L}_{\mathrm{MERA}} = \mathcal{L}_{\mathrm{shared}} + \mathcal{L}_{\mathrm{expert}}$.

## 3 Method

### 3.1 Overview of MedVTok

MedVTok learns a unified latent space that captures both structural fidelity and clinical semantics for 2D and 3D data via two complementary supervision mechanisms: (1) Slice Differential Regularization, which preserves volumetric anatomical continuity beyond what per-voxel reconstruction can capture (§3.2), and (2) Multi-Expert Representation Alignment (MERA), which integrates clinical knowledge from multiple modality-specific expert encoders into the shared token space (§3.3).

Fig. 2 illustrates the training framework. Specifically, given an input image or volume $\boldsymbol{x} \in \mathbb{R}^{C\times T\times H\times W}$, with $T{=}1$ for images and $T{>}1$ for volumes, MedVTok first partitions it into $N$ non-overlapping patches of size $C{\times}P_t{\times}P_h{\times}P_w$, where $N{=}\frac{THW}{P_tP_hP_w}$. Patches from different samples in a batch are joined into a sequence, where each patch is associated with a coordinate that records its position within the original input. This enables variable-resolution and mixed-dimension batching.

Then, the transformer encoder $E$, with rotary position embeddings (RoPE) [27] derived from the coordinates, maps the sequence to per-patch tokens $\boldsymbol{h} = E(\boldsymbol{x}) \in \mathbb{R}^{N\times d}$. These tokens are, on the one hand, regularized by a KL divergence term $\mathcal{L}_{\text{KL}}$ to impose a standard Gaussian constraint after projection and reparameterization; on the other hand, they are processed by the decoder $D$ to reconstruct the original input $\hat{\boldsymbol{x}}$. The reconstruction is supervised by the reconstruction loss $\mathcal{L}_{\text{R}}$ and the proposed slice differential regularization $\mathcal{L}_{\text{SliDR}}$. In parallel, an aggregation head $f$ aggregates the tokens $\boldsymbol{h}$ into a holistic representation $\boldsymbol{r} = f(\boldsymbol{h}) \in \mathbb{R}^d$, [1] which is used for semantic alignment via MERA.

### 3.2 Slice Differential Regularization

In tomographic volumes such as CT and MRI scans, adjacent slices share strong anatomical context: organs span continuously across the depth axis, and their boundaries shift gradually from slice to slice. Per-voxel reconstruction losses like $\ell_1$ can produce outputs that are individually accurate yet exhibit through-plane discontinuities, compromising diagnostic utility. We address this with the *Slice Differential Regularization*, which operates on inter-slice continuity rather than individual voxels.

For a 3D volume with $T{>}2$ slices along the depth axis, let $\boldsymbol{x}_{(k)}$ and $\hat{\boldsymbol{x}}_{(k)}$ denote the $k$-th axial slice of the input and the reconstruction, respectively. We define the first- and second-order slice differentials $\Delta_k(\boldsymbol{x}) = \boldsymbol{x}_{(k+1)} - \boldsymbol{x}_{(k)}$ and $\Delta_k^2(\boldsymbol{x}) = \boldsymbol{x}_{(k+1)} - 2\boldsymbol{x}_{(k)} + \boldsymbol{x}_{(k-1)}$, and penalize discrepancies between input and reconstructed transitions:

$$\mathcal{L}_{\text{SliDR}} = \frac{1}{T-1}\sum_{k=1}^{T-1} \left\|\Delta_k(\hat{\boldsymbol{x}}) - \Delta_k(\boldsymbol{x})\right\|_1 + \frac{1}{T-2}\sum_{k=2}^{T-1} \left\|\Delta_k^2(\hat{\boldsymbol{x}}) - \Delta_k^2(\boldsymbol{x})\right\|_1, \tag{1}$$

where the first term matches inter-slice transitions and the second preserves through-plane curvature.

### 3.3 Multi-Expert Representation Alignment

Apart from structural fidelity, a unified medical visual tokenizer should also encode high-level clinical semantics. Achieving this across the full breadth of medical imaging is challenging as different modalities embed distinct knowledge. Aligning to a single expert [6, 15] constrains the tokenizer to inherit a single semantic geometry, which may underrepresent modalities outside the expert's training distribution. We therefore propose *Multi-Expert Representation Alignment* (MERA), which uses a panel of pretrained modality-specific encoders as semantic *experts*. MERA involves two objectives: (i) shared-space contrastive alignment, which jointly aligns representations from multiple modality experts into a shared space, and (ii) expert-space complementary alignment, which preserves complementary information in each modality expert's representation space.

**Shared-space contrastive alignment.** An input $\boldsymbol{x}_i$ of imaging modality $m(i)$ is routed to the corresponding expert $E_{m(i)}$ to produce a feature $\boldsymbol{e}_i = E_{m(i)}(\boldsymbol{x}_i)$, which will be linearly projected to a shared space, giving $\tilde{\boldsymbol{e}}_i = p_{m(i)}(\boldsymbol{e}_i) \in \mathbb{R}^d$. We then apply a contrastive loss $\mathcal{L}_{\text{shared}}$ between MedVTok's representations $\boldsymbol{r}_i$ and expert features $\tilde{\boldsymbol{e}}_i$ within a batch with a learnable temperature $\tau$:

$$\mathcal{L}_{\text{shared}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(\boldsymbol{r}_i\cdot\tilde{\boldsymbol{e}}_i/\tau)}{\sum_{j=1}^{B}\exp(\boldsymbol{r}_i\cdot\tilde{\boldsymbol{e}}_j/\tau)} + \log\frac{\exp(\boldsymbol{r}_i\cdot\tilde{\boldsymbol{e}}_i/\tau)}{\sum_{j=1}^{B}\exp(\boldsymbol{r}_j\cdot\tilde{\boldsymbol{e}}_i/\tau)}\right]. \tag{2}$$

**Expert-space complementary alignment.** Note that $\mathcal{L}_{\text{shared}}$ projects heterogeneous expert features into a single space, which may discard modality-specific nuances present in each expert's native space. To address this issue, we introduce expert-space complementary learning. For each modality

[1] We find that, rather than implementing $f$ as the attention pooling used in prior work [9, 28], adopting simple mean pooling with a linear layer improves dense tasks while retaining performance on other tasks (§4.3).

$m(i)$, a backprojector $q_{m(i)}$ maps $\boldsymbol{r}_i$ back to the expert's space, yielding $\tilde{\boldsymbol{r}}_i = q_{m(i)}(\boldsymbol{r}_i) \in \mathbb{R}^{d_{m(i)}}$. We then enforce cosine similarity between backprojected representations and original expert features:

$$\mathcal{L}_{\text{expert}} = \frac{1}{B}\sum_{i=1}^{B}\left(1 - \frac{\tilde{\boldsymbol{r}}_i \cdot \boldsymbol{e}_i}{\|\tilde{\boldsymbol{r}}_i\| \cdot \|\boldsymbol{e}_i\|}\right). \tag{3}$$

Together, $\mathcal{L}_{\text{shared}}$ and $\mathcal{L}_{\text{expert}}$ provide bidirectional representation alignment: the former anchors MedVTok within a single representation space, while the latter preserves modality-specific expert knowledge that might otherwise be compressed away. The total loss for MERA is given by:

$$\mathcal{L}_{\text{MERA}} = \mathcal{L}_{\text{shared}} + \mathcal{L}_{\text{expert}}. \tag{4}$$

### 3.4 Training and Inference

MedVTok is trained with the combined objective:

$$\mathcal{L} = \mathcal{L}_{\text{R}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{SliDR}}\mathcal{L}_{\text{SliDR}} + \lambda_{\text{MERA}}\mathcal{L}_{\text{MERA}}, \tag{5}$$

where the standard [7, 10, 29] reconstruction loss $\mathcal{L}_{\text{R}}$ consists of pixel-level, LPIPS [30], and adversarial losses. Training proceeds in two stages: Stage 1 uses only 2D images to build foundational visual structure and semantic capabilities, with $\lambda_{\text{SliDR}} = 0$; Stage 2 continues on both 2D images and 3D volumes with $\lambda_{\text{SliDR}} = 0.1$. At inference, MedVTok serves as a general-purpose tokenizer: given an image or volume, it encodes tokens that are directly consumed by downstream models and transferable across various tasks without further fine-tuning of the tokenizer.

## 4 Experiments

### 4.1 Pretraining

**Training data.** We prepare the training data by aggregating existing large-scale publicly available training-ready datasets to ensure image quality and diversity. In total, we obtain 28,769,502 images and 70,144 volumes for training MedVTok. Please refer to § B.1 for data statistics.

**Implementation details.** We use the AdamW [31] optimizer with $(\beta_1, \beta_2) = (0.9, 0.95)$ and a cosine annealing schedule for 100K steps per stage. The learning rate decays from $10^{-4}$ to $10^{-5}$ in Stage 1 and from $2\times10^{-5}$ to $2\times10^{-6}$ in Stage 2, with $\lambda_{\text{KL}} = 10^{-6}$, $\lambda_{\text{MERA}} = 1$, and $\lambda_{\text{SliDR}} = 0.1$. During Stage 2 training, 2D and 3D data are fed into MedVTok in alternating iterations, with an effective batch size of 600 for 2D images and 48 for 3D volumes. Please see § B.2 for details.

### 4.2 Evaluation on Downstream Tasks

**Evaluation protocol.** We compare MedVTok with both medical-domain tokenizers [5, 6, 8] and general-domain ones [4, 9, 10, 29, 32]. We first evaluate token fidelity of different tokenizers via autoencoding reconstruction, and then evaluate token representation quality on 6 downstream tasks: image classification, image-to-image retrieval, image segmentation, image synthesis, visual question answering (VQA), and medical report generation. *All evaluations use a frozen tokenizer and focus on the quality of the encoded tokens rather than building task-specific models* (§ C.2).

**Evaluation data.** The token fidelity, classification, and retrieval evaluations share 9 benchmarks: JSIEC (fundus, retinal conditions) [33], ChestCTScan (CT, chest cancers) [34], HAM10K (dermoscopy, skin diseases) [35], BRISC (MRI, brain tumor) [36], PolypDiag (endoscopy, polyp detection) [37], OCTDL (OCT, retinal diseases) [38], LungHist700 (pathology, lung cancers) [39], BUSBRA (ultrasound, tumor staging) [40], and Kermany (X-ray, pneumonia) [41]. Segmentation datasets include MaplesDR (fundus, optic disc/cup) [42], MosMedDataPlus (CT, COVID-19 lesion) [43], HAM10K (dermoscopy, lesion) [35], PolypDB (endoscopy, polyp) [44], BRISC (MRI, brain tumor) [36], PanNuke (pathology, different cells) [45], BUSBRA (ultrasound, tumor) [40], and VinDr-RibCXR (X-ray, rib) [46]. Image synthesis uses images from the GMAI-VL-5.5M [1] dataset. VQA benchmarks include GMAI-MMBench [47], Fundus-MMBench [48], PathVQA [49], MM-Skin [50], Kvasir-VQA-x1 [51], and SLAKE [52]. For 3D tasks, we use AMOS-MM [53] and MRISegmenter [54] to evaluate 3D token fidelity. AMOS-MM CT–report pairs are used for report generation evaluation (§ C.1).

Table 1: **Token fidelity evaluation via autoencoding reconstruction** [PSNR (dB)↑, SSIM (%)↑]. 3D columns are excluded from the 2D average. MedVAE uses separate models for 2D and 3D inputs. The best result is marked in **bold** and the second-best is underlined.

| Metrics | Method | CFP | CT | Derm. | Endo. | MRI | OCT | Path. | US | X-ray | Avg. | 3D-MRI | 3D-CT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PSNR (↑) | SDVAE [29] | 37.62 | 36.38 | 34.68 | 33.57 | 34.31 | 29.17 | 24.43 | 32.13 | 35.29 | 33.06 | – | – |
| | TokenFlow [4] | 30.57 | 25.45 | 29.54 | 25.75 | 24.51 | 24.79 | 20.17 | 22.82 | 26.48 | 25.56 | – | – |
| | MingTok [32] | 35.58 | 31.23 | 32.43 | 29.85 | 29.37 | 26.92 | 21.87 | 26.42 | 31.25 | 29.43 | – | – |
| | UniTok [10] | 34.02 | 31.32 | 31.89 | 30.34 | 30.82 | 28.00 | 24.40 | 29.48 | 31.57 | 30.21 | – | – |
| | AToken [9] | 41.08 | 41.60 | 39.00 | 39.69 | 41.64 | 32.94 | 28.57 | 39.41 | 40.02 | 38.22 | 31.46 | 35.68 |
| | PUMIT [8] | 32.03 | 32.86 | 27.05 | 28.73 | 32.88 | 27.06 | 19.43 | 23.26 | 26.69 | 27.78 | 29.96 | 31.87 |
| | MedVAE [5] | 17.11 | 37.97 | 22.42 | 16.83 | 35.72 | 30.90 | 20.13 | 33.47 | 37.04 | 27.95 | 25.89 | 31.30 |
| | MedITok [6] | 35.92 | 35.61 | 33.81 | 32.42 | 34.82 | 30.26 | 26.03 | 32.62 | 34.70 | 32.91 | – | – |
| | MedVTok | **42.22** | **44.75** | **40.39** | **41.42** | **43.75** | **35.04** | **30.16** | **42.08** | **41.83** | **40.18** | **31.77** | **36.50** |
| SSIM (↑) | SDVAE [29] | 89.99 | 90.47 | 85.73 | 90.87 | 91.53 | 64.05 | 53.85 | 86.91 | 90.24 | 82.63 | – | – |
| | TokenFlow [4] | 83.75 | 73.40 | 78.03 | 77.23 | 74.13 | 59.89 | 31.75 | 55.34 | 76.99 | 67.84 | – | – |
| | MingTok [32] | 88.63 | 83.66 | 79.14 | 83.83 | 83.49 | 55.11 | 38.00 | 66.76 | 83.88 | 73.61 | – | – |
| | UniTok [10] | 88.54 | 88.70 | 82.48 | 89.62 | 90.12 | 65.24 | 55.36 | 87.21 | 88.20 | 81.72 | – | – |
| | AToken [9] | 94.74 | 96.24 | 92.51 | 96.98 | 97.61 | 80.60 | 74.24 | 97.13 | 95.42 | 91.72 | 85.19 | 95.50 |
| | PUMIT [8] | 88.09 | 90.82 | 84.84 | 90.17 | 90.92 | 65.34 | 57.92 | 81.99 | 89.30 | 82.16 | 85.20 | 90.55 |
| | MedVAE [5] | 77.26 | 92.67 | 86.63 | 83.86 | 93.87 | 72.44 | 54.18 | 89.97 | 92.90 | 82.64 | 75.14 | 90.44 |
| | MedITok [6] | 91.50 | 92.74 | 88.22 | 93.47 | 95.13 | 73.81 | 64.52 | 92.95 | 92.46 | 87.20 | – | – |
| | MedVTok | **95.95** | **97.86** | **94.51** | **97.90** | **98.66** | **86.54** | **80.52** | **98.28** | **97.21** | **94.16** | **87.79** | **95.65** |

**Token fidelity evaluation.** We assess token fidelity by how well the tokenizer reconstructs input data from its encoded tokens, reporting PSNR and SSIM between the input and the reconstructed data. Table 1 reports 2D reconstruction quality. MedVTok ranks first on all nine modalities in both PSNR and SSIM. The improvement is most notable in challenging domains such as pathology, where other tokenizers suffer from limited training diversity. Medical-domain tokenizers (MedITok, MedVAE) and general-domain ones (TokenFlow, UniTok, MingTok) lag significantly behind, confirming that medical images benefit from specialized architectures and large-scale multimodal training. Fig. 3 visualizes the reconstruction error maps of the competing methods and the reconstructed images produced by MedVTok. The last two columns of Table 1 report native 3D reconstruction. MedVTok achieves the best PSNR and SSIM on both MRI and CT volumes by a large margin over PUMIT and MedVAE. The strong 3D performance suggests that our unified 2D–3D training paradigm effectively transfers learned representations to volumetric data. Please see § C.2 for qualitative comparisons.

**Image classification and retrieval evaluation.** To directly measure the representation quality for image classification, we adopt a $k$-nearest neighbor protocol. Each test sample is classified based on its $k$ nearest training neighbors with $k \in \{10, 20, 100, 200\}$, and we report the highest accuracy among $k$ values for each tokenizer. For retrieval, the training split serves as the gallery, and the test split provides the queries. Gallery images are ranked by cosine similarity, and we report Recall@1 performance.

Results are presented in Table 2. MedVTok achieves the highest average on both tasks, surpassing not only other tokenizers but also dedicated semantic encoders (BiomedCLIP, MedImageInsight, MedSigLIP) that are explicitly trained for representation learning. This indicates that the features produced by MedVTok are highly discriminative. The consistent ranking further confirms that the learned representations generalize across different downstream protocols.

**Image segmentation evaluation.** We follow Siméoni *et al.* [55] and adopt a linear segmentation protocol to evaluate the representation quality for image segmentation. Specifically, a single linear

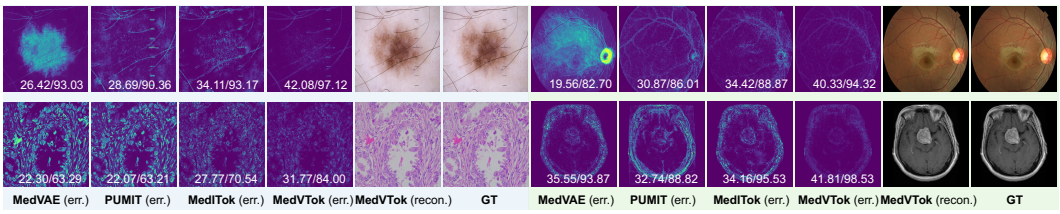



Figure 3: **Qualitative reconstruction for token fidelity evaluation.** For each example, we show per-tokenizer absolute error maps with PSNR/SSIM values reported per panel.

Table 2: **Image classification and retrieval evaluation** [Accuracy (%)↑, Recall@1 (%)↑]. *: semantic encoder.

| Metrics | Method | CFP | CT | Derm. | Endo. | MRI | OCT | Path. | US | X-ray | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Acc. (↑) | SDVAE [29] | 18.79 | 45.68 | 60.23 | 47.59 | 77.47 | 62.94 | 47.56 | 43.97 | 78.69 | 53.66 |
| | TokenFlow [4] | 35.73 | 46.10 | 61.95 | 42.83 | 79.73 | 68.53 | 45.38 | 46.10 | 86.91 | 57.03 |
| | MingTok [32] | 50.71 | 48.10 | 67.86 | 44.93 | 77.70 | 64.39 | 64.87 | 45.04 | 83.65 | 60.81 |
| | UniTok [10] | 66.95 | 74.79 | 70.63 | 44.82 | 80.53 | 72.88 | 69.23 | 46.28 | 91.03 | 68.57 |
| | AToken [9] | 64.12 | 65.82 | 72.31 | 46.63 | 88.13 | 66.46 | 66.54 | 46.19 | 91.29 | 67.50 |
| | BiomedCLIP* [15] | 60.59 | 66.03 | 68.78 | 46.03 | 88.43 | 73.91 | 79.10 | 48.32 | 93.80 | 69.45 |
| | MedSigLIP* [17] | 81.07 | 73.73 | 72.13 | 46.11 | 91.73 | 75.88 | **84.10** | 46.99 | 95.30 | 74.12 |
| | MedImageInsight* [16] | 75.64 | 75.32 | **85.87** | 44.65 | **95.47** | 90.79 | 71.79 | 50.18 | 96.21 | 76.21 |
| | PUMIT [8] | 13.84 | 43.78 | 60.34 | 45.85 | 78.00 | 59.52 | 47.44 | 43.44 | 79.49 | 52.41 |
| | MedVAE [5] | 13.28 | 45.25 | 60.12 | 45.34 | 75.57 | 57.56 | 49.62 | 41.76 | 73.40 | 51.32 |
| | MedITok [6] | 74.29 | 78.80 | 72.82 | 46.80 | 91.23 | 78.67 | 74.23 | 50.09 | 95.03 | 73.55 |
| | MedVTok | **90.25** | **89.77** | 74.16 | **48.71** | 94.23 | **94.00** | 82.69 | **51.33** | **96.90** | **80.23** |
| Recall@1 (↑) | SDVAE [29] | 22.75 | 43.81 | 57.41 | **56.14** | 90.10 | 57.14 | 32.30 | 39.10 | 82.53 | 53.47 |
| | TokenFlow [4] | 45.06 | 50.79 | 61.38 | 45.08 | 91.00 | 68.63 | 35.02 | 44.41 | 87.50 | 58.76 |
| | MingTok [32] | 54.94 | 52.38 | 63.16 | 44.82 | 89.50 | 68.01 | 53.70 | 42.29 | 83.01 | 61.31 |
| | UniTok [10] | 69.10 | 88.25 | 64.22 | 50.48 | 89.80 | 73.60 | 67.70 | 44.87 | 87.98 | 70.67 |
| | AToken [9] | 66.09 | 78.10 | 68.72 | 44.04 | 95.10 | 74.53 | 63.42 | 43.62 | 88.46 | 69.12 |
| | BiomedCLIP* [15] | 62.66 | 83.81 | 66.27 | 44.56 | 96.00 | 76.40 | 70.43 | 45.48 | 89.10 | 70.52 |
| | MedSigLIP* [17] | 82.40 | **92.06** | 67.79 | 44.82 | 96.90 | 77.02 | **83.27** | 39.63 | 94.55 | 75.38 |
| | MedImageInsight* [16] | 78.54 | 86.67 | **84.13** | 44.30 | **97.80** | 92.86 | 70.04 | 46.54 | **95.19** | 77.34 |
| | PUMIT [8] | 15.88 | 49.52 | 58.66 | 48.96 | 89.80 | 51.24 | 38.52 | 36.70 | 83.49 | 52.53 |
| | MedVAE [5] | 12.45 | 47.30 | 56.15 | 53.69 | 88.90 | 52.48 | 38.52 | 35.90 | 76.28 | 51.30 |
| | MedITok [6] | 73.82 | 87.62 | 68.65 | 52.52 | 95.60 | 80.75 | 72.76 | 46.01 | 92.15 | 74.43 |
| | MedVTok | **90.99** | 90.16 | 70.70 | 52.59 | 96.10 | **93.79** | 75.49 | **47.61** | **95.19** | **79.18** |

Table 3: **Image segmentation evaluation** [Dice (%)↑]. *: semantic encoder.

| Metrics | Method | CFP | CT | Derm. | Endo. | MRI | Path. | US | X-ray | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| Dice (↑) | SDVAE [29] | 73.03 | 14.88 | 62.37 | 24.67 | 12.69 | 49.11 | 84.19 | 35.56 | 44.56 |
| | TokenFlow [4] | 70.39 | 19.05 | 72.62 | 40.53 | 30.49 | 52.34 | 91.33 | 49.52 | 53.28 |
| | MingTok [32] | 85.36 | 27.03 | 83.91 | 57.32 | 52.54 | 67.68 | 92.91 | 54.71 | 65.18 |
| | UniTok [10] | 60.93 | 33.13 | 86.25 | 67.25 | 54.90 | 72.51 | 93.35 | 57.60 | 65.74 |
| | AToken [9] | 97.09 | 37.33 | 87.45 | 69.88 | 63.57 | 74.30 | 94.90 | 59.69 | 73.03 |
| | BiomedCLIP* [15] | 50.22 | 31.83 | 87.48 | 60.64 | 57.95 | 73.65 | 95.41 | 59.83 | 64.63 |
| | MedSigLIP* [17] | 97.28 | 40.79 | 87.76 | 69.81 | 67.32 | 76.20 | 95.33 | **62.48** | 74.62 |
| | MedImageInsight* [16] | 92.34 | 25.60 | 85.45 | 66.96 | 55.81 | 42.93 | 95.80 | 50.16 | 64.38 |
| | PUMIT [8] | 40.05 | 25.53 | 77.94 | 37.32 | 36.69 | 66.00 | 90.73 | 54.82 | 53.63 |
| | MedVAE [5] | 61.74 | 12.57 | 57.59 | 19.73 | 10.13 | 36.40 | 81.90 | 30.81 | 38.86 |
| | MedITok [6] | 96.70 | 40.52 | 80.28 | 66.07 | 62.42 | 75.80 | 94.65 | 59.47 | 71.99 |
| | MedVTok | **97.44** | **42.91** | **90.25** | **72.52** | **69.62** | **77.74** | **96.91** | 61.13 | **76.07** |

segmentation head is trained with a cross-entropy loss on top of the frozen patch tokens produced by each tokenizer, with learning rates searched over $\{10^{-2}, 10^{-3}, 10^{-4}\}$. For each tokenizer, the best-performing head on the validation set is selected for evaluation. Results are presented in Table 3. MedVTok achieves the best average Dice score, outperforming all baselines including semantic encoders. This task reveals a clear hierarchy: although general autoencoding tokenizers (*e.g.*, SDVAE, PUMIT) can embed rich low-level details into the latent space, their high-level semantics are not activated, resulting in low segmentation performance. Semantic encoders provide explicit visual semantics and significantly outperform the autoencoding tokenizers. Unified tokenizers (MedITok, MedVTok) encode both fine-grained spatial information and high-level semantics, enjoying the benefits of both and improving dense prediction. Please see § C.2 for qualitative comparisons.

**Image synthesis evaluation.** We evaluate how the tokenizer influences text-conditioned medical image synthesis within a latent diffusion [29] framework. A UNet denoising backbone is trained from scratch in each tokenizer's latent space while keeping both the tokenizer and the text encoder frozen. This design isolates the effect of the tokenizer: the architecture, noise scheduler, optimizer, and training data are held constant across experiments. Training uses 70K image–text pairs spanning seven imaging modalities. We report the Fréchet Inception Distance (FID) [56, 57] in Table 4; lower FID indicates higher fidelity. MedVTok achieves the highest image synthesis quality with the best average FID, outperforming both specialized tokenizers (PUMIT, MedVAE) and the previous

Table 4: **Image synthesis evaluation** (FID ↓). A latent diffusion model is trained from scratch in each frozen tokenizer's latent space with identical architecture and data.

| Method | CFP | CT | Derm. | Endo. | MRI | Path. | X-ray | Avg. |
|---|---|---|---|---|---|---|---|---|
| PUMIT [8] | 140.40 | 183.49 | 272.26 | 191.49 | 227.64 | 377.12 | 200.14 | 227.51 |
| MedVAE [5] | 118.13 | 154.42 | 202.09 | 170.29 | 145.25 | 325.87 | 88.55 | 172.08 |
| MedITok [6] | 90.65 | **142.46** | 176.94 | **58.38** | 122.96 | 99.66 | 77.24 | 109.76 |
| MedVTok | **59.85** | 146.39 | **142.87** | 82.39 | **116.78** | **94.65** | **69.17** | **101.73** |

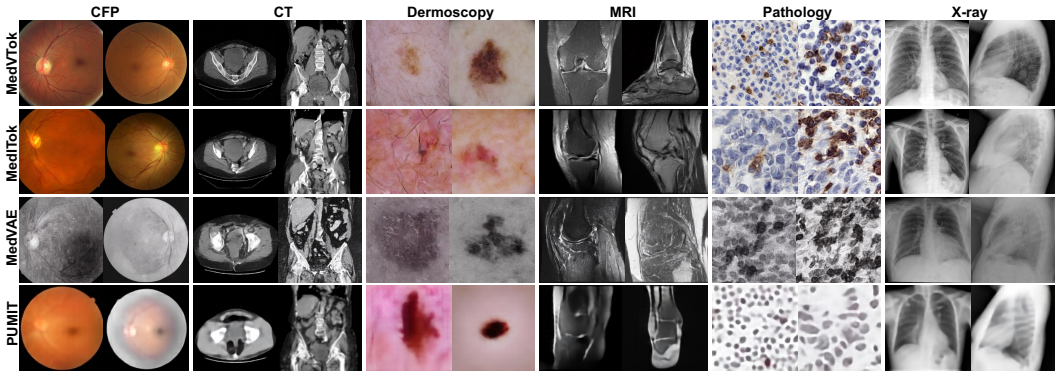



Figure 4: **Image synthesis examples.** A latent diffusion model trained from scratch on MedVTok's tokens produces samples with sharper anatomical structure and fewer artifacts than competing tokenizers across modalities, including fundus, X-ray, and dermoscopy.

unified tokenizer (MedITok), as shown in Fig. 4. PUMIT and MedVAE consistently underperform, confirming that synthesis quality depends not only on reconstruction fidelity but also on a well-structured, semantically meaningful latent space that facilitates the diffusion model's learning.

Table 5: **Visual question answering evaluation** [Accuracy (%)↑]. Each frozen tokenizer is plugged into the LLaVA-Med [58] framework for evaluation.

| Method | FundusMM | MM-Skin | Kvasir | PathVQA | SLAKE | GMAI-MM | Avg. |
|---|---|---|---|---|---|---|---|
| PUMIT [8] | 29.84±0.42 | 34.74±0.38 | 69.50±0.51 | 52.97±0.63 | 36.44±0.55 | 28.16±0.47 | 41.94±0.33 |
| MedVAE [5] | 28.39±0.56 | 34.21±0.44 | 66.64±0.72 | 51.15±0.58 | 31.57±0.67 | 27.61±0.53 | 39.93±0.41 |
| MedITok [6] | 31.29±0.35 | 36.18±0.41 | 70.63±0.29 | **58.46**±0.52 | 45.61±0.38 | 30.15±0.26 | 45.39±0.28 |
| MedVTok | **33.39**±0.31 | **37.49**±0.28 | **72.98**±0.34 | 57.24±0.45 | **45.94**±0.37 | **31.81**±0.29 | **46.48**±0.22 |

**Visual question answering evaluation.** We evaluate each tokenizer as a *frozen* visual backbone within the LLaVA-Med [58] framework, training only the LLaVA-Med connector and the LLM, and evaluating the resulting models on VQA benchmarks [6]. We report accuracy assessed via LLM-as-judge grading. As shown in Table 5, MedVTok achieves the best average accuracy among the competing tokenizers. Notably, MedVTok does not use image–text pairs for training, yet still outperforms MedITok, which relies on millions of textual captions for representation learning.

**Medical report generation evaluation.** Evaluation on generating free-text radiology reports from 3D volumes is similar to VQA, incorporating each frozen tokenizer into the LaMed framework [59] and training the connector and LLM to generate region-specific findings for three anatomical regions (abdomen, chest, and pelvis). We report BLEU-2, RaTEScore [60] (entity-level NER-based evaluation), and GREEN [61] (clinically aware error detection via an LLM judge). Table 6 presents the report generation results. MedVTok achieves the best scores on all metrics. The improvement is consistent across anatomical regions, with particularly strong gains in the pelvic region, which is the most challenging due to its complex anatomy and longer reference reports. Compared with MedITok, MedVTok shows meaningful improvements on clinically aware metrics (GREEN: 0.413 vs. 0.386), suggesting that our tokenizer encodes diagnostically relevant information that translates into more accurate clinical language generation. Additional results are provided in § C.2.

Table 6: **Medical report generation evaluation.** Each frozen tokenizer is integrated into the LaMed [59] framework for evaluation, generating findings in abdomen, chest, and pelvis regions.

| Metrics | Method | Abdomen | Chest | Pelvis | Avg. |
|---|---|---|---|---|---|
| BLEU-2 (↑) | PUMIT [8] | **0.289**±0.008 | 0.193±0.003 | 0.164±0.007 | 0.215±0.005 |
| | MedVAE-3D [5] | 0.249±0.038 | 0.185±0.020 | 0.152±0.010 | 0.195±0.013 |
| | MedITok [6] | 0.256±0.013 | 0.219±0.020 | **0.294**±0.017 | 0.257±0.012 |
| | MedVTok | 0.272±0.005 | **0.263**±0.045 | 0.288±0.017 | **0.274**±0.009 |
| RaTEScore (↑) | PUMIT [8] | 0.604±0.005 | 0.511±0.035 | 0.201±0.015 | 0.439±0.006 |
| | MedVAE-3D [5] | 0.613±0.022 | 0.497±0.053 | 0.187±0.003 | 0.432±0.011 |
| | MedITok [6] | 0.610±0.006 | 0.501±0.008 | 0.594±0.031 | 0.568±0.015 |
| | MedVTok | **0.619**±0.005 | **0.570**±0.074 | **0.612**±0.006 | **0.600**±0.006 |
| GREEN (↑) | PUMIT [8] | 0.421±0.017 | 0.204±0.010 | 0.124±0.019 | 0.250±0.009 |
| | MedVAE-3D [5] | 0.401±0.024 | 0.186±0.003 | 0.105±0.064 | 0.230±0.179 |
| | MedITok [6] | 0.412±0.025 | 0.256±0.010 | 0.492±0.026 | 0.386±0.013 |
| | MedVTok | **0.425**±0.005 | **0.300**±0.078 | **0.515**±0.044 | **0.413**±0.020 |

### 4.3 Ablation Studies

We ablate the key components of MedVTok in Fig. 5. For efficiency, most components are evaluated on a training subset (3M images) in Stage 1 prior to data scaling.

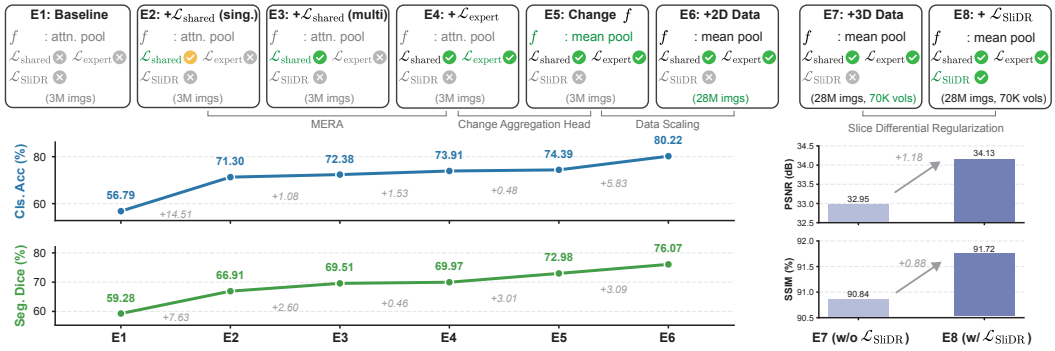



Figure 5: **Ablation studies.**

**Effect of MERA components.** Starting from a baseline without any semantic alignment (E1 in Fig. 5), shared-space contrastive learning with a single-expert $\mathcal{L}_{\text{shared}}$ (E2, using MedImageInsight [16]) yields an initial improvement in classification and segmentation, confirming that semantic supervision is critical for downstream understanding. Extending $\mathcal{L}_{\text{shared}}$ to multiple experts (E3) further improves both metrics, demonstrating that a panel of modality-specific experts provides more comprehensive semantics. Adding expert-space complementary learning $\mathcal{L}_{\text{expert}}$ (E4) brings an additional +1.53% classification gain by preserving modality-specific nuances that shared-space contrastive learning alone may compress away during the shared-space projection.

**Aggregation head design.** Replacing attention pooling [9, 28] with mean pooling (E4→E5) yields substantial gains in segmentation. Attention pooling tends to collapse patch-level details into a global summary, whereas mean pooling retains spatially distributed information critical for dense prediction tasks while still achieving competitive classification performance.

**Data scaling.** Scaling the training set from 3M to 28M images (E5→E6) yields consistent improvements, confirming that our framework effectively leverages more diverse training data.

**Effect of SliDR.** Stage 2 (E7) adds 3D data to enable volumetric tokenization. Adding $\mathcal{L}_{\text{SliDR}}$ (E8) improves 3D reconstruction in terms of both PSNR and SSIM, validating that SliDR specifically strengthens through-plane continuity. Please see § C.3 for additional analyses.

## 5 Conclusion

We present MedVTok, a general-purpose medical visual tokenizer that encodes both 2D images and 3D volumes into a shared representational space through a single transformer architecture. Two complementary supervision mechanisms shape this space at different levels: Slice Differential Regularization preserves inter-slice fidelity that per-voxel supervision misses, and Multi-Expert Representation Alignment injects modality-specific clinical semantics from a panel of experts without requiring paired annotations. MedVTok achieves state-of-the-art results across reconstruction and six downstream tasks, ranging from pixel-level fidelity to report generation, demonstrating that structural fidelity and semantic richness can be jointly optimized within a single tokenizer.

We acknowledge several limitations. First, the training data focus on common imaging modalities at present. Other imaging modalities, clinical photos, or illustrations in medical publications remain unexplored. Incorporating these would enhance generalization to non-standard visual inputs encountered in real-world clinical workflows. Second, while MERA uses frozen expert encoders, jointly fine-tuning selected experts or incorporating new ones could further improve domain coverage when the computational budget allows. We hope MedVTok will serve as a useful building block for scalable multimodal medical AI systems that enable patient-level modeling across heterogeneous imaging evidence and clinical tasks.

## Acknowledgments and Disclosure of Funding

This work was supported in part by National Natural Science Foundation of China (Nos. 62471148 and T2596013), STI2030-Major Projects (No. 2021ZD0200204), the 2025 High-Quality Industrial Development Special Program for Scientific Intelligence ("Baituan Baixiang," Second Batch; No. 2025-GZL-RGZN-BTBX-02015), and Shanghai Center for Brain Science and Brain-inspired Technology. The authors declare no competing interests.

## References


[1] Tianbin Li, Yanzhou Su, Wei Li, Bin Fu, Zhe Chen, Ziyan Huang, Guoan Wang, Chenglong Ma, Ying Chen, Ming Hu, et al. GMAI-VL & GMAI-VL-5.5M: A large vision-language model and a comprehensive multimodal dataset towards general medical AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(28):23177–23185, Mar. 2026.

[2] Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Song xiaohui, et al. HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=WbP2OwMULq.

[3] Junzhi Ning, Wei Li, Cheng Tang, Jiashi Lin, Chenglong Ma, Chaoyang Zhang, Jiyao Liu, Ying Chen, Shujian Gao, Lihao Liu, et al. UnimedVL: Unifying medical multimodal understanding and generation through observation-knowledge-analysis. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=7UR3rNXMz9.

[4] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. TokenFlow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2545–2555, 2025.

[5] Maya Varma, Ashwin Kumar, Rogier Van der Sluijs, Sophie Ostmeier, Louis Blankemeier, Pierre Joseph Marcel Chambon, Christian Bluethgen, Jip Prince, Curtis Langlotz, and Akshay S Chaudhari. MedVAE: Efficient automated interpretation of medical images with large-scale generalizable autoencoders. In *Medical Imaging with Deep Learning*, 2025. URL https://openreview.net/forum?id=jNmvKHniMe.

[6] Chenglong Ma, Yuanfeng Ji, Jin Ye, Zilong Li, Chenhui Wang, Junzhi Ning, Wei Li, Lihao Liu, Qiushan Guo, Tianbin Li, Junjun He, and Hongming Shan. MedITok: A unified tokenizer for medical image synthesis and interpretation. *arXiv preprint arXiv:2505.19225*, 2025.

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021.

[8] Lingxiao Luo, Xuanzhong Chen, Bingda Tang, Xinsheng Chen, Rong Han, Chengpeng Hu, Yujiang Li, and Ting Chen. Building universal foundation models for medical image analysis with spatially adaptive networks. *arXiv preprint arXiv:2312.07630*, 2023.

[9] Jiasen Lu, Liangchen Song, Mingze Xu, Byeongjoo Ahn, Yanjun Wang, Chen Chen, Afshin Dehghan, and Yinfei Yang. Atoken: A unified tokenizer for vision. *arXiv preprint arXiv:2509.14476*, 2025.

[10] Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. UniTok: A unified tokenizer for visual generation and understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=f6aOPkGE8L.

[11] Shilong Zhang, He Zhang, Zhifei Zhang, Chongjian GE, Shuchen Xue, Shaoteng Liu, Mengwei Ren, Soo Ye Kim, Yuqian Zhou, Qing Liu, Daniil Pakhomov, Kai Zhang, Zhe Lin, and Ping Luo. Both semantics and reconstruction matter: Making representation encoders ready for text-to-image generation and editing. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=UmlXZ4t44h.

[12] Aaron Van Den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[13] Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *Advances in Neural Information Processing Systems*, 37:28281–28295, 2024.

[14] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications*, 16(1):7866, 2025.

[15] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640, 2025. doi: 10.1056/AIoa2400640.

[16] Noel CF Codella, Ying Jin, Shrey Jain, Yu Gu, Ho Hin Lee, Asma Ben Abacha, Alberto Santamaria-Pang, Will Guyman, Naiteek Sangani, Sheng Zhang, et al. Medimageinsight: An open-source embedding model for general domain medical imaging. *arXiv preprint arXiv:2410.06542*, 2024.

[17] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[18] Min Woo Sun, Alejandro Lozano, Javier Gamazo Tejero, Vishwesh Nath, Xiao Xiao Sun, James Burgess, Yuhui Zhang, Kun Yuan, Robert Tibshirani, Sean Huver, et al. No tokens wasted: Leveraging long context in biomedical vision-language models. *arXiv preprint arXiv:2510.03978*, 2025.

[19] Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Unimed-CLIP: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. *arXiv preprint arXiv:2412.10372*, 2024.

[20] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024.

[21] Siyuan Yan, Ming Hu, Yiwen Jiang, Xieji Li, Hao Fei, Philipp Tschandl, Harald Kittler, and Zongyuan Ge. Derm1m: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12681–12690, 2025.

[22] Jiawei Du, Jia Guo, Weihang Zhang, Shengzhu Yang, Hanruo Liu, Huiqi Li, and Ningli Wang. RET-CLIP: A retinal image foundation model pre-trained with clinical diagnostic reports. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15012. Springer Nature Switzerland, October 2024.

[23] Meng Wang, Tian Lin, Aidi Lin, Kai Yu, Yuanyuan Peng, Lianyu Wang, Cheng Chen, Ke Zou, Huiyu Liang, Man Chen, et al. Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model. *Nature Communications*, 16(1):5528, 2025.

[24] Kun yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=zuwpeRkJNH.

[25] Cris Claessens, Christiaan Viviers, Giacomo D'Amicantonio, Egor Bondarev, and Fons van der Sommen. Scaling self-supervised and cross-modal pretraining for volumetric CT transformers. *arXiv preprint arXiv:2511.17209*, 2025.

[26] Chenhui Zhao, Yiwei Lyu, Asadur Zaman Chowdury, Edward S Harake, Akhil Kondepudi, Akshay T Rao, Xinhai Hou, Honglak Lee, and Todd C Hollon. Towards scalable language-image pre-training for 3d medical imaging. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=WxHf4EcBWA.

[27] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[28] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv:2502.14786*, 2025.

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[30] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[32] Ziyuan Huang, DanDan Zheng, Cheng Zou, Rui Liu, Xiaolong Wang, Kaixiang Ji, Weilong Chai, Jianxin Sun, Libin Wang, Yongjie Lv, et al. Ming-univision: Joint image understanding and generation with a unified continuous tokenizer. *arXiv preprint arXiv:2510.06590*, 2025.

[33] Ling-Ping Cen, Jie Ji, Jian-Wei Lin, Si-Tong Ju, Hong-Jie Lin, Tai-Ping Li, Yun Wang, Jian-Feng Yang, Yu-Fen Liu, Shaoying Tan, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature communications*, 12(1): 4828, 2021.

[34] Sibtain Syed, Rehan Ahmed, and Arshad Iqbal. Medical imagining (ct scan, mri, x-ray, and microscopic imagery) data, 2024. URL https://doi.org/10.17632/5kbjrgsncf.3.

[35] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5 (1):180161, 2018.

[36] Amirreza Fateh, Yasin Rezvani, Sara Moayedi, Sadjad Rezvani, Fatemeh Fateh, and Mansoor Fateh. Brisc: Annotated dataset for brain tumor segmentation and classification with swin-hafnet. *arXiv e-prints*, pages arXiv–2506, 2025.

[37] Yu Tian, Guansong Pang, Fengbei Liu, Yuyuan Liu, Chong Wang, Yuanhong Chen, Johan Verjans, and Gustavo Carneiro. Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2022.

[38] Mikhail Kulyabin, Aleksei Zhdanov, Anastasia Nikiforova, Andrey Stepichev, Anna Kuznetsova, Mikhail Ronkin, Vasilii Borisov, Alexander Bogachev, Sergey Korotkich, Paul A Constable, et al. Octdl: Optical coherence tomography dataset for image-based deep learning methods. *Scientific data*, 11(1):365, 2024.

[39] Jorge Diosdado, Pere Gilabert, Santi Seguí, and Henar Borrego. Lunghist700: A dataset of histological images for deep learning in pulmonary pathology. *Scientific Data*, 11(1):1088, 2024.

[40] Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. Bus-bra: A breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical physics*, 51(4):3110–3123, 2024.

[41] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

[42] Gabriel Lepetit-Aimon, Clément Playout, Marie Carole Boucher, Renaud Duval, Michael H Brent, and Farida Cheriet. Maples-dr: Messidor anatomical and pathological labels for explainable screening of diabetic retinopathy. *Scientific Data*, 11(1):914, 2024.

[43] Sergey P Morozov, Anna E Andreychenko, Nikolay A Pavlov, AV Vladzymyrskyy, Natalya V Ledikhova, Victor A Gombolevskiy, Ivan A Blokhin, Pavel B Gelezhe, AV Gonchar, and V Yu Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*, 2020.

[44] Debesh Jha, Nikhil Kumar Tomar, Vanshali Sharma, Quoc-Huy Trinh, Koushik Biswas, Hongyi Pan, Ritika K Jha, Gorkem Durak, Alexander Hann, Jonas Varkey, et al. Polypdb: A curated multi-center dataset for development of ai algorithms in colonoscopy. *arXiv preprint arXiv:2409.00045*, 2024.

[45] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *European congress on digital pathology*, pages 11–19. Springer, 2019.

[46] Hoang C Nguyen, Tung T Le, Hieu H Pham, and Ha Q Nguyen. Vindr-ribcxr: A benchmark dataset for automatic segmentation and labeling of individual ribs on chest x-rays. *arXiv preprint arXiv:2107.01327*, 2021.

[47] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024.

[48] Xinyao Liu and Diping Song. Constructing ophthalmic mllm for positioning-diagnosis collaboration through clinical cognitive chain reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21547–21556, 2025.

[49] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[50] Wenqi Zeng, Yuqi Sun, Chenxi Ma, Weimin Tan, and Bo Yan. Mm-skin: Enhancing dermatology vision-language model with an image-text dataset derived from textbooks. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3769–3778, 2025.

[51] Sushant Gautam, Michael Riegler, and Pål Halvorsen. Kvasir-vqa-x1: A multimodal dataset for medical reasoning and robust medvqa in gastrointestinal endoscopy. In *MICCAI Workshop on Data Engineering in Medical Imaging*, pages 53–63. Springer, 2025.

[52] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[53] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022.

[54] Yan Zhuang, Tejas Sudharshan Mathai, Pritam Mukherjee, Brandon Khoury, Boah Kim, Benjamin Hou, Nusrat Rabbee, Abhinav Suri, and Ronald M Summers. MRISegmenter: A fully accurate and robust automated multiorgan and structure segmentation tool for t1-weighted abdominal mri. *Radiology*, 315(1):e241979, 2025.

[55] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

[56] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

[57] McKell Woodland, Austin Castelo, Mais Al Taie, Jessica Albuquerque Marques Silva, Mohamed Eltaher, Frank Mohn, Alexander Shieh, Suprateek Kundu, Joshua P. Yung, Ankit B. Patel, and Kristy K. Brock. Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 87–97, 2024.

[58] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=GSuP99u2kR.

[59] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3D: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

[60] Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. RaTEScore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, 2024.

[61] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, November 2024.

[62] Yikang Zhou, Tao Zhang, Dengxian Gong, Yuanzheng Wu, Ye Tian, Haochen Wang, Haobo Yuan, Jiacong Wang, Lu Qi, Hao Fei, Shunping Ji, Anran Wang, Zhuochen Wang, Yujing Wang, Cheng Chen, and Xiangtai Li. SAMTok: Representing any mask with two words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 37852–37863, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhou_SAMTok_Representing_Any_Mask_with_Two_Words_CVPR_2026_paper.html.

[63] Tao Wang, Changxu Cheng, Lingfeng Wang, Senda Chen, and Wuyue Zhao. HiMTok: Learning hierarchical mask tokens for image segmentation with large multimodal model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23267–23278, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_HiMTok_Learning_Hierarchical_Mask_Tokens_for_Image_Segmentation_with_Large_ICCV_2025_paper.html.

[64] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

[65] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in Neural Information Processing Systems*, 29, 2016.

[66] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. HuatuoGPT-Vision, towards injecting medical visual knowledge into multimodal LLMs at scale. *arXiv:2406.19280*, 2024.

[67] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[68] Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30:105474, 2020.

[69] MD Anouk Stein, Carol Wu, Chris Carr, Errol Colak, George Shih, Jeff Rudie, John Mongan, Julia Elliott, Luciano Prevedello, MD Marc Kohli, Phil Culliton, and Robyn Ball. RSNA STR Pulmonary Embolism Detection, 2020. URL https://www.kaggle.com/competitions/rsna-str-pulmonary-embolism-detection. Kaggle.

[70] Shuang Song, Rui Xu, Yong Luo, Bo Du, Jiancheng Yang, Kaiming Kuang, Yunlang She, and Mengmeng Zhao. Mela dataset: A benchmark for mediastinal lesion analysis (training set part 1), May 2022. URL https://doi.org/10.5281/zenodo.6575197.

[71] Liang Jin, Jiancheng Yang, Kaiming Kuang, Bingbing Ni, Yiyi Gao, Yingli Sun, Pan Gao, Weiling Ma, Mingyu Tan, Hui Kang, Jiajun Chen, and Ming Li. Deep-learning-assisted detection and segmentation of rib fractures from ct scans: Development and validation of fracnet. *eBioMedicine*, 2020.

[72] Minghui Zhang, Yangqian Wu, Hanxiao Zhang, Yulei Qin, Hao Zheng, Wen Tang, Corey Arnold, Chenhao Pei, Pengxin Yu, Yang Nan, et al. Multi-site, multi-domain airway tree modeling. *Medical Image Analysis*, 90:102957, 2023.

[73] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2):915–931, 2011.

[74] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Mae, Adamo Young, Cheng Zhu, Xin Yang, Kangkang Meng, Ziyan Huang, et al. Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the FLARE22 challenge. *The Lancet Digital Health*, 6(11):e815–e826, 2024.

[75] R. R. Beichel, E. J. Ulrich, C. Bauer, A. Wahle, B. Brown, T. Chang, K. Plichta, B. Smith, J. Sunderland, T. Braun, A. Fedorov, D. Clunie, M. Onken, V. A. Magnotta, Y. Menda, J. Riesmeier, S. Pieper, R. Kikinis, M. M. Graham, T. L. Casavant, M. Sonka, and J. Buatti. Data from qin-headneck (version 4) [data set], 2015. URL https://www.cancerimagingarchive.net/collection/qin-headneck.

[76] Petr Jordan, Philip M Adamson, Vrunda Bhattbhatt, Surabhi Beriwal, Sangyu Shen, Oskar Radermecker, Supratik Bose, Linda S Strain, Michael Offe, David Fraley, et al. Pediatric chest-abdomen-pelvis and abdomen-pelvis CT images with expert organ contours. *Medical Physics*, 49(5):3523–3528, 2022.

[77] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. AbdomenCT-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6695–6714, 2021.

[78] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

[79] Mitchell Machtay, Fenghai Duan, Barry A Siegel, Bradley S Snyder, Jeremy J Gorelick, Janet S Reddin, Reginald Munden, Douglas W Johnson, Larry H Wilf, Albert DeNittis, et al. Prediction of survival by [18f] fluorodeoxyglucose positron emission tomography in patients with locally advanced non–small-cell lung cancer undergoing definitive chemoradiation therapy: results of the acrin 6668/rtog 0235 trial. *Journal of clinical oncology*, 31(30):3823–3830, 2013.

[80] J. Saltz, M. Saltz, P. Prasanna, R. Moffitt, J. Hajagos, E. Bremer, J. Balsamo, and T. Kurc. Stony brook university covid-19 positive cases [data set], 2021. URL https://www.cancerimagingarchive.net/collection/covid-19-ny-sbu.

[81] Gongning Luo, Kuanquan Wang, Jun Liu, Shuo Li, Xinjie Liang, Xiangyu Li, Shaowei Gan, Wei Wang, Suyu Dong, Wenyi Wang, et al. Efficient automatic segmentation for multi-level pulmonary arteries: The parse challenge. *arXiv preprint arXiv:2304.03708*, 2023.

[82] Pengbo Liu, Hu Han, Yuanqi Du, Heqin Zhu, Yinhao Li, Feng Gu, Honghu Xiao, Jun Li, Chunpeng Zhao, Li Xiao, Xinbao Wu, and S. Kevin Zhou. Deep learning to segment pelvic bones: large-scale ct datasets and baseline models. *International Journal of Computer Assisted Radiology and Surgery*, 16(5):749, 2021.

[83] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (LiTS). *Medical Image Analysis*, 84:102680, 2023.

[84] Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601, 2022.

[85] National Lung Screening Trial Research Team. The national lung screening trial: overview and study design. *Radiology*, 258(1):243–253, 2011.

[86] Germán González, Daniel Jimenez-Carretero, Sara Rodríguez-López, Carlos Cano-Espinosa, Miguel Cazorla, Tanya Agarwal, Vinit Agarwal, Nima Tajbakhsh, Michael B Gotway, Jianming Liang, et al. Computer aided detection for pulmonary embolism challenge (cad-pe). *arXiv preprint arXiv:2003.13440*, 2020.

[87] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022.

[88] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):381, 2020.

[89] João Pedrosa, Guilherme Aresta, Carlos Ferreira, Márcio Rodrigues, Patrícia Leitão, André Silva Carvalho, João Rebelo, Eduardo Negrão, Isabel Ramos, António Cunha, et al. LNDb: a lung nodule database on computed tomography. *arXiv preprint arXiv:1911.08434*, 2019.

[90] Aaron J Grossberg, Abdallah SR Mohamed, Hesham Elhalawani, William C Bennett, Kirk E Smith, Tracy S Nolan, Bowman Williams, Sasikarn Chamchod, Jolien Heukelom, Michael E Kantor, Theodora Browne, , Katherine A Hutcheson, G Brandon Gunn, Adam S Garden, William H Morrison, Steven J Frank, David I Rosenthal, John B Freymann, and Clifton D Fuller. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. *Scientific Data*, 5(1):1–10, 2018.

[91] Bram Van Ginneken, Samuel G Armato III, Bartjan de Hoop, Saskia van Amelsvoort-van de Vorst, Thomas Duindam, Meindert Niemeijer, Keelin Murphy, Arnold Schilham, Alessandra Retico, Maria Evelina Fantacci, et al. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the anode09 study. *Medical Image Analysis*, 14(6):707–722, 2010.

[92] Yang Deng, Ce Wang, Yuan Hui, Qian Li, Jun Li, Shiwei Luo, Mengke Sun, Quan Quan, Shuxin Yang, You Hao, et al. Ctspine1k: A large-scale dataset for spinal vertebrae segmentation in computed tomography. *arXiv preprint arXiv:2105.14711*, 2021.

[93] National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium head and neck squamous cell carcinoma collection (cptac-hnscc) (version 19) [data set], 2018. URL https://www.cancerimagingarchive.net/collection/cptac-hnscc.

[94] P. Li, S. Wang, T. Li, J. Lu, Y. HuangFu, and D. Wang. A large-scale ct and pet/ct dataset for lung cancer diagnosis (Lung-PET-CT-Dx) [data set], 2020. URL https://www.cancerimagingarchive.net/collection/lung-pet-ct-dx.

[95] P. Kinahan, M. Muzi, B. Bialecki, B. Herman, and L. Coombs. Data from the acrin 6668 trial nsclc-fdg-pet (version 2) [data set], 2019. URL https://www.cancerimagingarchive.net/collection/acrin-nsclc-fdg-pet.

[96] W. R. Bosch, W. L. Straube, J. W. Matthews, and J. A. Purdy. Head-neck cetuximab [data set], 2015. URL https://www.cancerimagingarchive.net/collection/head-neck-cetuximab.

[97] P. Kinahan, M. Muzi, B. Bialecki, and L. Coombs. Data from acrin-flt-breast (version 2) [data set], 2017. URL https://www.cancerimagingarchive.net/collection/acrin-flt-breast.

[98] Stephanie A Harmon, Thomas H Sanford, Sheng Xu, Evrim B Turkbey, Holger Roth, Ziyue Xu, Dong Yang, Andriy Myronenko, Victoria Anderson, Amel Amalou, et al. Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. *Nature Communications*, 11(1):4080, 2020.

[99] M. Patnana, S. Patel, and A. Tsao. Anti-PD-1 immunotherapy melanoma dataset [data set], 2019.

[100] Ahmed Shahin, Carmela Wegworth, David, Elizabeth Estes, Julia Elliott, Justin Zita, SimonWalsh, Slepetys, and Will Cukierski. Osic pulmonary fibrosis progression. https://kaggle.com/competitions/osic-pulmonary-fibrosis-progression, 2020. Kaggle.

[101] Samuel G. Armato, Lubomir Hadjiiski, Georgia D. Tourassi, Karen Drukker, Maryellen L. Giger, Feng Li, George Redmond, Keyvan Farahani, Justin S. Kirby, and Laurence P. Clarke. Spie-aapm-nci lung nodule classification challenge dataset, 2015. URL https://doi.org/10.7937/K9/TCIA.2015.UZLSU3FL.

[102] Emily B Tsai, Scott Simpson, Matthew P Lungren, Michelle Hershman, Leonid Roshkovan, Errol Colak, Bradley J Erickson, George Shih, Anouk Stein, Jayashree Kalpathy-Cramer, et al. The rsna international covid-19 open radiology database (ricord). *Radiology*, 299(1): E204–E213, 2021.

[103] B. Albertina, M. Watson, C. Holback, R. Jarosz, S. Kirk, Y. Lee, K. Rieger-Christ, and J. Lemmerman. The cancer genome atlas lung adenocarcinoma collection (tcga-luad) (version 4) [data set], 2016. URL https://www.cancerimagingarchive.net/collection/tcga-luad.

[104] P. Madhavi, S. Patel, and A. S. Tsao. Data from anti-pd-1 immunotherapy lung [data set], 2019. URL https://www.cancerimagingarchive.net/collection/anti-pd-1_lung.

[105] National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium lung adenocarcinoma collection (cptac-luad) (version 12) [data set], 2018. URL https://www.cancerimagingarchive.net/collection/cptac-luad.

[106] S. Desai, A. Baghal, T. Wongsurawat, S. Al-Shukri, K. Gates, P. Farmer, M. Rutherford, G. D. Blake, T. Nolan, T. Powell, et al. Data from chest imaging with clinical and genomic correlates representing a rural covid-19 positive population, 2020. URL https://doi.org/10.7937/tcia.2020.py71-5978. [Data set].

[107] Joel Saltz, Michael Saltz, Prateek Prasanna, Richard Moffitt, Justin Hajagos, Emily Bremer, Joseph Balsamo, and Tahsin Kurc. Stony brook university covid-19 positive cases, 2021. URL https://doi.org/10.7937/TCIA.BBAG-2923. [Data set].

[108] S. Kirk, Y. Lee, P. Kumar, J. Filippini, B. Albertina, M. Watson, K. Rieger-Christ, and J. Lemmerman. The cancer genome atlas lung squamous cell carcinoma collection (tcga-lusc), 2016. URL https://doi.org/10.7937/K9/TCIA.2016.TYGKKFMQ. Version 4 [Data set].

[109] Xiangde Luo, Jia Fu, Yunxin Zhong, Shuolin Liu, Bing Han, Mehdi Astaraki, Simone Bendazzoli, Iuliana Toma-Dasu, Yiwen Ye, Ziyang Chen, et al. SegRap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma. *Medical image analysis*, 101:103447, 2025.

[110] Anjany Sekuboyina, Malek E. Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, Martin Urschler, Maodong Chen, Dalong Cheng, Nikolas Lessmann, Yujin Hu, Tianfu Wang, Dong Yang, Daguang Xu, Felix Ambellan, Tamaz Amiranashvili, Moritz Ehlke, Hans Lamecker, Sebastian Lehnert, Marilia Lirio, Nicolás Pérez de Olaguer, Heiko Ramm, Manish Sahu, Alexander Tack, Stefan Zachow, Tao Jiang, Xinjun Ma, Christoph Angerman, Xin Wang, Kevin Brown, Alexandre Kirszenberg, Élodie Puybareau, Di Chen, Yiwei Bai, Brandon H. Rapazzo, Timyoas Yeah, Amber Zhang, Shangliang Xu, Feng Hou, Zhiqiang He, Chan Zeng, Zheng Xiangshang, Xu Liming, Tucker J. Netherton, Raymond P. Mumme, Laurence E. Court, Zixun Huang,

Chenhang He, Li-Wen Wang, Sai Ho Ling, Lê Duy Huỳnh, Nicolas Boutry, Roman Jakubicek, Jiri Chmelik, Supriti Mulay, Mohanasankar Sivaprakasam, Johannes C. Paetzold, Suprosanna Shit, Ivan Ezhov, Benedikt Wiestler, Ben Glocker, Alexander Valentinitsch, Markus Rempfler, Björn H. Menze, and Jan S. Kirschke. Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images. *Medical Image Analysis*, 73:102166, 2021.

[111] Nicholas Heller, Niranjan Sathianathen, Anurag Kalapara, Eric Walczak, Kelly Moore, Hanna Kaluzniak, Justin Rosenberg, Peter Blake, Zachary Rengel, Michael Oestreich, John Dean, Michael Tradewell, Ankit Shah, Rohan Tejpaul, Zachary Edgerton, Matthew Peterson, Saad Raza, Saugat Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. Data from c4kc-kits, 2019. URL https://doi.org/10.7937/TCIA.2019.IX49E8NX. [Data set].

[112] Bram van Ginneken. Lola11, April 2021. URL https://doi.org/10.5281/zenodo.4708800.

[113] Pechin Lo, Bram Van Ginneken, Joseph M Reinhardt, Tarunashree Yavarna, Pim A De Jong, Benjamin Irving, Catalin Fetita, Margarete Ortner, Rômulo Pinho, Jan Sijbers, et al. Extraction of airways from CT (EXACT'09). *IEEE Transactions on Medical Imaging*, 31(11):2093–2107, 2012.

[114] Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, et al. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. *arXiv preprint arXiv:2307.01984*, 2023.

[115] Benjamin Movsas, Chen Hu, Jeffrey Sloan, Jeffrey Bradley, Ritsuko Komaki, Gregory Masters, Vivek Kavadi, Samir Narayan, Jeff Michalski, Douglas W. Johnson, Christopher Koprowski, Jr Curran, Walter J., Yolanda I. Garces, Rakesh Gaur, Raymond B. Wynn, John Schallenkamp, Daphna Y. Gelblum, Robert M. MacRae, Rebecca Paulus, and Hak Choy. Quality of Life Analysis of a Radiation Dose–Escalation Study of Patients With Non–Small-Cell Lung Cancer: A Secondary Analysis of the Radiation Therapy Oncology Group 0617 Randomized Clinical Trial. *JAMA Oncology*, 2(3):359–367, 03 2016.

[116] Onur Akin, Paul Elnajjar, Michael Heller, Richard Jarosz, Bradley J. Erickson, Scott Kirk, Y. Lee, W. Marston Linehan, R. Gautam, Rohini Vikram, et al. The cancer genome atlas kidney renal clear cell carcinoma collection (tcga-kirc), 2016. URL https://doi.org/10.7937/K9/TCIA.2016.V6PBVTDR. Version 3 [Data set].

[117] National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium cutaneous melanoma collection (cptac-cm), 2018. URL https://doi.org/10.7937/K9/TCIA.2018.ODU24GZE. Version 11 [Data set].

[118] Rina D. Rudyanto, Sjoerd Kerkstra, Eva M. van Rikxoort, Catalin Fetita, Pierre-Yves Brillet, Christophe Lefevre, Wenzhe Xue, Xiangjun Zhu, Jianming Liang, İlkay Öksüz, Devrim Ünay, Kamuran Kadipaşaog˘lu, Raúl San José Estépar, James C. Ross, George R. Washko, Juan-Carlos Prieto, Marcela Hernández Hoyos, Maciej Orkisz, Hans Meine, Markus Hüllebrand, Christina Stöcker, Fernando Lopez Mir, Valery Naranjo, Eliseo Villanueva, Marius Staring, Changyan Xiao, Berend C. Stoel, Anna Fabijanska, Erik Smistad, Anne C. Elster, Frank Lindseth, Amir Hossein Foruzan, Ryan Kiros, Karteek Popuri, Dana Cobzas, Daniel Jimenez-Carretero, Andres Santos, Maria J. Ledesma-Carbayo, Michael Helmberger, Martin Urschler, Michael Pienn, Dennis G.H. Bosboom, Arantza Campo, Mathias Prokop, Pim A. de Jong, Carlos Ortiz de Solorzano, Arrate Muñoz-Barrutia, and Bram van Ginneken. Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: the vessel12 study. *Medical Image Analysis*, 18(7):1217–1232, 2014.

[119] Bradley J. Erickson, David Mutch, Laura Lippmann, and Richard Jarosz. The cancer genome atlas uterine corpus endometrial carcinoma collection (tcga-ucec), 2016. URL https://doi.org/10.7937/K9/TCIA.2016.GKJ0ZWAC. Version 4 [Data set].

[120] C. Holback, R. Jarosz, F. Prior, D. G. Mutch, P. Bhosale, K. Garcia, Y. Lee, S. Kirk, C. A. Sadow, S. Levine, et al. The cancer genome atlas ovarian cancer collection (tcga-ov), 2016. URL https://doi.org/10.7937/K9/TCIA.2016.NDO1MDFQ. Version 4 [Data set].

[121] J. Y. Y. Kwan, J. Su, S. H. Huang, L. S. Ghoraie, W. Xu, B. Chan, K. W. Yip, M. Giuliani, A. Bayley, J. Kim, A. J. Hope, J. Ringash, J. Cho, A. McNiven, A. Hansen, D. Goldstein, J. R. de Almeida, H. J. Aerts, J. N. Waldron, B. Haibe-Kains, B. O'Sullivan, S. V. Bratman, and F. F. Liu. Data from radiomic biomarkers to refine risk models for distant metastasis in oropharyngeal carcinoma [data set], 2019. URL https://www.cancerimagingarchive.net/collection/opc-radiomics.

[122] National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium sarcomas collection (cptac-sar), 2019. URL https://doi.org/10.7937/tcia.2019.9bt23r95. Version 10 [Data set].

[123] Holger R. Roth, Amal Farag, Evrim B. Turkbey, Le Lu, Jiamin Liu, and Ronald M. Summers. Data from pancreas-ct. https://doi.org/10.7937/K9/TCIA.2016.tNB1kqBU, 2016. The Cancer Imaging Archive.

[124] Michael Green, Edith M Marom, Eli Konen, Nahum Kiryati, and Arnaldo Mayer. Feature aggregation in perceptual loss for ultra low-dose (uld) ct denoising. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1635–1638, 2019.

[125] P. Kinahan, M. Muzi, B. Bialecki, and L. Coombs. Ddata from the acrin 6685 trial hnscc-fdg-pet/ct [data set], 2019. URL https://www.cancerimagingarchive.net/collection/acrin-hnscc-fdg-pet-ct.

[126] Holger R Roth, Ziyue Xu, Carlos Tor-Díez, Ramon Sanchez Jacob, Jonathan Zember, Jose Molto, Wenqi Li, Sheng Xu, Baris Turkbey, Evrim Turkbey, et al. Rapid artificial intelligence solutions in a pandemic—the covid-19-20 lung ct lesion segmentation challenge. *Medical Image Analysis*, 82:102605, 2022.

[127] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315, 2022.

[128] Nicholas Kurtansky, Veronica Rotemberg, Maura Gillis, Kivanc Kose, Walter Reade, and Ashley Chow. Isic 2024: Skin cancer detection with 3d-tbp. https://kaggle.com/competitions/isic-2024-challenge, 2024. Kaggle competition.

[129] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8(1):34, 2021.

[130] Seung Seog Han, Ik Jun Moon, Seong Hwan Kim, Jung-Im Na, Myoung Shin Kim, Gyeong Hun Park, Ilwoo Park, Keewon Kim, Woohyung Lim, Ju Hee Lee, et al. Assessment of deep neural networks for the diagnosis of benign and malignant skin neoplasms in comparison with dermatologists: A retrospective validation study. *PLoS medicine*, 17(11):e1003381, 2020.

[131] Carlos Hernández-Pérez, Marc Combalia, Sebastian Podlipnik, Noel CF Codella, Veronica Rotemberg, Allan C Halpern, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Brian Helba, et al. BCN20000: Dermoscopic lesions in the wild. *Scientific Data*, 11(1):641, 2024.

[132] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828, 2021.

[133] Abdurrahim Yilmaz, Sirin Pekcan Yasar, Gulsum Gencoglan, and Burak Temelkuran. Derm12345: A large, multisource dermatoscopic skin lesion dataset with 40 subclasses. *Scientific Data*, 11(1):1302, 2024.

[134] Philipp Tschandl, Bengu Nisa Akay, Cliff Rosendahl, Veronica Rotemberg, Verche Todorovska, Jochen Weber, Anna Katharina Wolber, Christoph Müller, Nicholas Kurtansky, Allan Halpern,

Wolfgang Weninger, and Harald Kittler. MILK10k: A hierarchical multimodal imaging-learning toolkit for diagnosing pigmented and nonpigmented skin cancer and its simulators. *Journal of Investigative Dermatology*, 146(2):357–364.e7, 2026. ISSN 0022-202X.

[135] Ronald de Jong, HJ Carolus, HA Franciscus, Romy C van Jaarsveld, Richard van Hillegersberg, PW Josien, Peter HN de With, Yasmina al Khalil, Fons van Der Sommen, et al. Scaling up self-supervised learning for improved surgical foundation models. *Medical Image Analysis*, page 103873, 2025.

[136] Aneeq Zia, Max Berniker, Rogerio Garcia Nespolo, Conor Perreault, Kiran Bhattacharyya, Xi Liu, Ziheng Wang, Satoshi Kondo, Satoshi Kasai, Kousuke Hirasawa, Bo Liu, David Austin, Yiheng Wang, Michal Futrega, Jean-Francois Puget, Zhenqiang Li, Yoichi Sato, Ryo Fujii, Ryo Hachiuma, Mana Masuda, Hideo Saito, An Wang, Mengya Xu, Mobarakol Islam, Long Bai, Winnie Pang, Hongliang Ren, Chinedu Nwoye, Luca Sestini, Nicolas Padoy, Maximilian Nielsen, Samuel Schüttler, Thilo Sentker, Hümeyra Husseini, Ivo Baltruschat, Rüdiger Schmitz, René Werner, Aleksandr Matsun, Mugariya Farooq, Numan Saaed, Jose Renato Restom Viera, Mohammad Yaqub, Neil Getty, Fangfang Xia, Zixuan Zhao, Xiaotian Duan, Xing Yao, Ange Lou, Hao Yang, Jintong Han, Jack Noble, Jie Ying Wu, Tamer Abdulbaki Alshirbaji, Nour Aldeen Jalal, Herag Arabian, Ning Ding, Knut Moeller, Weiliang Chen, Quan He, Muhammad Bilal, Taofeek Akinosho, Adnan Qayyum, Massimo Caputo, Hunaid Vohra, Michael Loizou, Anuoluwapo Ajayi, Ilhem Berrou, Faatihah Niyi-Odumosu, Charlie Budd, Oluwatosin Alabi, Tom Vercauteren, Ruoxi Zhao, Ayberk Acar, John Han, Jumanh Atoum, Yinhong Qin, Jie Ying Wu, Surong Hua, Lu Ping, Wenming Wu, Rongfeng Wei, Jinlin Wu, You Pang, Zhen Chen, Tim Jaspers, Amine Yamlahi, Piotr Kalinowski, Dominik Michael, Tim Rä dsch, Marco Hübner, Danail Stoyanov, Stefanie Speidel, Lena Maier-Hein, and Anthony Jarc. Intuitive surgical surgtoolloc challenge results: 2022-2023. *arXiv preprint arXiv:2305.07152*, 2025.

[137] Vivek Singh Bawa, Gurkirt Singh, Francis KapingA, Inna Skarga-Bandurova, Elettra Oleari, Alice Leporini, Carmela Landolfo, Pengfei Zhao, Xi Xiang, Gongning Luo, et al. The SARAS endoscopic surgeon action detection (esad) dataset: Challenges and methods, 2021.

[138] Stefanie Speidel, Lena Maier-Hein, Danail Stoyanov, Sebastian Bodenstedt, Annika Reinke, Sophia Bano, Alexander Jenke, Martin Wagner, Marie Daum, Ala Tabibian, et al. Endoscopic vision challenge 2023. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2023)*. Zenodo, 2023. doi: 10.5281/zenodo.8315050. URL https://doi.org/10.5281/zenodo.8315050.

[139] Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, et al. AIROGS: Artificial intelligence for robust glaucoma screening challenge. *IEEE Transactions on Medical Imaging*, 43(1):542–557, 2024.

[140] Emma Dugas, Jared, Jorge, and Will Cukierski. Diabetic retinopathy detection. https://kaggle.com/competitions/diabetic-retinopathy-detection, 2015. Kaggle.

[141] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410, 2016.

[142] Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, et al. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19724–19735, 2025.

[143] Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. HISTAI: an open-source, large-scale whole slide image dataset for computational pathology. *arXiv preprint arXiv:2505.12120*, 2025.

[144] Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Zhongyi Shui, Kai Zhang, Jingxiong Li, Xingheng Lyu, Tao Lin, and Lin Yang. PathGen-1.6M: 1.6 million pathology image-text pairs generation through multi-agent collaboration. *arXiv preprint arXiv:2407.00203*, 2024.

[145] Christian Matek, Sebastian Krappe, Christian Münzenmayer, Torsten Haferlach, and Carsten Marr. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood, The Journal of the American Society of Hematology*, 138(20):1917–1927, 2021.

[146] Yiping Jiao, Jeroen van der Laak, Shadi Albarqouni, Zhang Li, Tao Tan, Abhir Bhalerao, Shenghua Cheng, Jiabo Ma, Johnathan Pocock, Josien P.W. Pluim, et al. Lysto: The lymphocyte assessment hackathon and benchmark dataset. *IEEE Journal of Biomedical and Health Informatics*, 28(3):1161–1172, 2024. doi: 10.1109/JBHI.2023.3327489.

[147] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[148] IXI. IXI Dataset. https://brain-development.org/ixi-dataset/, 2010. Imperial College London, Creative Commons CC BY-SA 3.0.

[149] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[150] Alessa Hering, Lasse Hansen, Tony CW Mok, Albert CS Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, et al. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 42(3):697–712, 2022.

[151] Jonathan Shapey, Aaron Kujawa, Reuben Dorent, Guotai Wang, Alexis Dimitriadis, Diana Grishchuk, Ian Paddick, Neil Kitchen, Robert Bradford, Shakeel R Saeed, et al. Segmentation of vestibular schwannoma from mri, an open annotated dataset and baseline algorithm. *Scientific Data*, 8(1):286, 2021.

[152] Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, 22(12):2677–2684, 2010.

[153] Sook-Lei Liew, Bethany P Lo, Miranda R Donnelly, Artemis Zavaliangos-Petropulu, Jessica N Jeong, Giuseppe Barisano, Alexandre Hutton, Julia P Simon, Julia M Juliano, Anisha Suri, et al. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific Data*, 9(1):320, 2022.

[154] Natalia Shusharina and Thomas Bortfeld. Glioma image segmentation for radiotherapy: Rt targets, barriers to cancer spread, and organs at risk (glis-rt), 2021. URL https://doi.org/10.7937/TCIA.T905-ZQ20. [Data set].

[155] Reuben Dorent, Aaron Kujawa, Marina Ivory, Spyridon Bakas, Nicola Rieke, Samuel Joutard, Ben Glocker, Jorge Cardoso, Marc Modat, Kayhan Batmanghelich, Arseniy Belkov, Maria Baldeon Calisto, Jae Won Choi, Benoit M. Dawant, Hexin Dong, Sergio Escalera, Yubo Fan, Lasse Hansen, Mattias P. Heinrich, Smriti Joshi, Victoriya Kashtanova, Hyeon Gyu Kim, Satoshi Kondo, Christian N. Kruse, Susana K. Lai-Yuen, Hao Li, Han Liu, Buntheng Ly, Ipek Oguz, Hyungseob Shin, Boris Shirokikh, Zixian Su, Guotai Wang, Jianghao Wu, Yanwu Xu, Kai Yao, Li Zhang, Sébastien Ourselin, Jonathan Shapey, and Tom Vercauteren. Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. *Medical Image Analysis*, 83:102628, 2023.

[156] Huijun Chen, Xihai Zhao, Haozhong Sun, Jiaqi Dou, Chenlin Du, Runyu Yang, Xiaoqi Lin, Han Jiang, Shuwan Yu, Jiachen Liu, Zixuan Han, Chun Yuan, and Niranjan Balu. Cerebral artery segmentation challenge (cas) 2023 [data set], 2023. URL https://codalab.lisn.upsaclay.fr/competitions/9804. Accessed: 2025-05-22.

[157] Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical image analysis*, 67:101832, 2021.

[158] Mohit Prabhushankar, Kiran Kokilepersaud, Yash-yee Logan, Stephanie Trejo Corona, Ghassan AlRegib, and Charles Wykoff. Olives dataset: Ophthalmic labels for investigating visual eye semantics. *Advances in Neural Information Processing Systems*, 35:9201–9216, 2022.

[159] Daniel Kermany. Labeled optical coherence tomography (OCT) and chest X-ray images for classification. *Mendeley data*, 2018.

[160] Weiyi Zhang, Peranut Chotcomwongse, Yinwen Li, Pusheng Xu, Ruijie Yao, Lianhao Zhou, Yuxuan Zhou, Hui Feng, Qiping Zhou, Xinyue Wang, et al. Predicting diabetic macular edema treatment responses using OCT: Dataset and methods of APTOS competition. *Medical Image Analysis*, 109:103942, 2026. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2026.103942.

[161] Malliga Subramanian, Kogilavani Shanmugavadivel, Obuli Sai Naren, K Premkumar, and K Rankish. Classification of retinal oct images using deep learning. In *2022 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–7, 2022. doi: 10.1109/ICCCI54379.2022.9740985.

[162] María Sofía Sappia. Acouslic-ai : Abdominal circumference operator- agnostic ultrasound measurement in low-income countries using artificial intelligence, July 2024. URL https://doi.org/10.5281/zenodo.12697994.

[163] Sun Yat-sen University. Sysu-ceus-fll: Contrast-enhanced ultrasound dataset for focal liver lesions. https://www.kaggle.com/datasets/hoangviet2302/sysu-ceus-fll, 2023. Kaggle dataset.

[164] Hanae Elmekki, Ahmed Alagha, Hani Sami, Amanda Spilkin, Antonela Mariel Zanuttini, Ehsan Zakeri, Jamal Bentahar, Lyes Kadem, Wen-Fang Xie, Philippe Pibarot, et al. Cactus: An open dataset and framework for automated cardiac assessment and classification of ultrasound images using deep transfer learning. *Computers in Biology and Medicine*, 190:110003, 2025.

[165] Hari M. Trivedi, Maryam Vazirabad, Felipe C. Kitamura, Yan Chen, and Helen Frazer. Open-source dataset for the rsna screening mammography cancer detection challenge. *Radiology: Artificial Intelligence*, 8(2):e250375, 2026.

[166] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):170177, 2017.

[167] Hieu T Nguyen, Ha Q Nguyen, Hieu H Pham, Khanh Lam, Linh T Le, Minh Dao, and Van Vu. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1):277, 2023.

[168] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

[169] Xiaoman Zhang, Julián N Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. ReXGradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. *arXiv preprint arXiv:2505.00228*, 2025.

[170] Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. CheXpert Plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024.

[171] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

[172] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2097–2106, 2017.

[173] Yifan Wu, Hayden Gunraj, Chi-en Amy Tai, and Alexander Wong. COVIDx CXR-4: An expanded multi-institutional open-source benchmark dataset for chest x-ray image-based computer-aided covid-19 diagnostics. *arXiv preprint arXiv:2311.17677*, 2023.

[174] Eduardo P Reis, Joselisa PQ De Paiva, Maria CB Da Silva, Guilherme AS Ribeiro, Victor F Paiva, Lucas Bulgarelli, Henrique MH Lee, Paulo V Santos, Vanessa M Brito, Lucas TW Amaral, et al. Brax, brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):487, 2022.

[175] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[176] Radiological Society of North America (RSNA). Rsna pneumonia detection challenge. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge, 2018. Chest X-ray dataset with pneumonia annotations.

[177] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

[178] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pages 1–19, 2026.

[179] Martin Vallieres, Emily Kay-Rivest, Léo Jean Perrin, Xavier Liem, Christophe Furstoss, Hugo JWL Aerts, Nader Khaouam, Phuc Felix Nguyen-Tan, Chang-Shu Wang, Khalil Sultanem, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Scientific Reports*, 7(1):10117, 2017.

[180] Forithmus. MR-RATE: A vision-language dataset for magnetic resonance imaging, 2026. URL https://huggingface.co/datasets/Forithmus/MR-RATE. Hugging Face.

[181] Meng Lou, Xiaoqing Liu, Yuqing Zhang, Yizhou Yu, and Hong-Yu Zhou. Liver lesion diagnosis challenge on multi-phase MRI, April 2023. URL https://doi.org/10.5281/zenodo.7852363.

[182] Carlos Martín-Isla, Víctor M Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, Alireza Sojoudi, Mitchell J Fulton, Tewodros Weldebirhan Arega, Kumaradevan Punithakumar, et al. Deep learning segmentation of the right ventricle in cardiac mri: the m&ms challenge. *IEEE Journal of Biomedical and Health Informatics*, 27(7): 3302–3313, 2023.

# Appendix

This appendix provides full experimental details and supplementary results. §A presents the tokenizer architecture, including input patchification, attention masking, positional embeddings, and expert encoders. §B describes the training settings: §B.1 details the training data construction and lists the 2D and 3D datasets; §B.2 specifies optimization, adversarial training, and the two-stage training procedure. §C covers evaluation: §C.1 summarizes all evaluation benchmarks grouped by downstream task, §C.2 details the evaluation protocols for each task along with additional quantitative and qualitative results, and §C.3 provides additional ablation studies and analyses. §D discusses broader societal impacts.

## A Tokenizer Overview

**Relation to existing tokenizers.** MedITok [6] combines reconstruction and semantic alignment for 2D medical images. AToken [9] provides a unified tokenizer for general-domain images, videos, and 3D assets. SAMTok [62] and HiMTok [63] encode segmentation masks as compact token sequences for multimodal language models. MedVTok focuses on a shared representation for medical images and volumes, combining depth-wise reconstruction regularization with semantic supervision from modality-specific experts.

**Architecture hyperparameters.** The encoder and decoder of MedVTok each consist of 27 transformer blocks with a hidden dimension of $d = 1152$. The encoder is initialized from SigLIP2-SO400M-patch16-NaFlex [28], while other components are trained from scratch. The patch size is $P_t \times P_h \times P_w$, where $P_t = 4$ is the depth patch size for 3D volumes and $P_h = 16, P_w = 16$ are the spatial patch sizes for both 2D images and 3D volumes. Rotary position embeddings (RoPE) [27] are derived from 3D patch coordinates $(z, x, y)$. Before the decoder, each encoded $d$-dimensional token embedding is projected to a 96-dimensional output, which is split into mean and log-variance for a diagonal Gaussian posterior, yielding latent codes of dimension $d_z = 48$.

**Dimension-agnostic input patchification.** All inputs, regardless of spatial resolution, aspect ratio, or dimensionality, are converted into a common *concatenated patch* format before entering the transformer encoder. Each patch spans $P_t \times P_h \times P_w = 4 \times 16 \times 16$ voxels across the depth, height, and width axes, yielding a per-patch feature vector of dimension $P_t \cdot P_h \cdot P_w \cdot C = 3072$ for $C = 3$ channels. For a 2D image of spatial size $H \times W$, the input is first zero-padded along the temporal axis to $P_t$ images. Patchification then yields $(H/P_h) \times (W/P_w) = \frac{HW}{256}$ patches (*e.g.*, a $256 \times 256$ image produces 256 patches). For a 3D volume of depth $T$ and spatial size $H \times W$, patchification directly operates along all three axes, producing $(T/P_t) \times (H/P_h) \times (W/P_w)$ patches (*e.g.*, a 32-slice $256 \times 256$ CT volume yields $8 \times 16 \times 16 = 2048$ patches).

All patches from a batch are concatenated into a single flat sequence of shape $[N, 3072]$, where $N$ is the total number of patches across the batch. This eliminates dense padding entirely and achieves $\mathcal{O}(N)$ memory regardless of resolution heterogeneity. Each patch is associated with a coordinate vector $(b, z, x, y)$ encoding the batch index, depth position, and spatial position, together with a list of per-sample sequence lengths that enables $\mathcal{O}(1)$ batch-wise slicing. Because the representation has no global spatial grid, a single batch can freely mix different spatial resolutions (*e.g.*, $256 \times 256$ and $512 \times 512$), different depths (*e.g.*, 16 and 64 slices), and 2D images with 3D volumes simultaneously.

**Attention masking for concatenated sequences.** Self-attention in the transformer blocks operates over the full concatenated sequence; however, cross-sample information leakage is prevented via block-diagonal attention masks, restricting each token to attend only within the boundaries of its own sample.

**Expert encoders for MERA.** Table 7 lists the expert encoders used in MERA. Features from all experts are precomputed offline and stored alongside the training data, so the expert encoders themselves are not loaded during training. Indices 9–10 (3D experts) are included only in Stage 2.

Each input uses the expert assigned to its modality, with a dedicated projector and backprojector. Shared-space alignment brings these representations into one token space, while expert-space alignment retains modality-specific information. Complementary knowledge thus refers to the distinct semantics contributed by experts across modalities.

Table 7: **Expert encoders used in MERA.**

| Index | Expert Encoder | Domains | $d_j$ |
|---|---|---|---|
| 0 | DermLIP [21] | Dermoscopy | 768 |
| 1 | RetCLIP [22] | CFP | 768 |
| 2 | Conch [20] | Pathology | 1024 |
| 3 | PeskaVLP [24] | Endoscopy | 2048 |
| 4–8 | MedImageInsight [16] | CT, MR, US, X-ray, Other | 2048 |
| 9 | Spectre [25] | CT 3D | 2160 |
| 10 | HLIP [26] | MR 3D | 2048 |

# B Training Settings

## B.1 Training Datasets

Figs. 14 and 15 provide a visual overview of the training data composition; Tables 20 and 21 list the 2D and 3D training datasets, respectively. In total, we collect 28,769,502 images and 70,144 volumes spanning 9 imaging modalities.

We curate the training set from publicly available medical imaging repositories. Before quality filtering, we impose minimum dataset size thresholds: only 2D datasets containing at least 9,000 images and 3D datasets containing at least 180 volumes are included. This criterion serves two purposes. First, it prevents commonly used small-scale benchmark datasets from entering the training pool, thereby preserving their integrity for downstream evaluation. Second, it avoids highly fragmented data sources that contribute minimal training signal relative to the overhead of integration.

Following MedITok [6], we apply a series of automated quality filters to remove low-quality or clinically irrelevant 2D images, including checks on dynamic range, minimum resolution, pixel-value diversity, and information content, as well as manual screening to exclude non-clinical illustrations such as tables and plots.

## B.2 Implementation Details

As stated in the main text (Eq. 5), the total training objective is $\mathcal{L} = \mathcal{L}_{\text{R}} + \lambda_{\text{SliDR}}\mathcal{L}_{\text{SliDR}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{MERA}}\mathcal{L}_{\text{MERA}}$. Below we detail the loss components, the two-stage training procedure, and the optimization settings.

**Loss functions.** The reconstruction loss is defined as:

$$\mathcal{L}_{\text{R}} = \ell_1 + \ell_2 + \mathcal{L}_{\text{LPIPS}} + w_{\text{adv}}\mathcal{L}_{\text{adv}}, \tag{6}$$

where $\ell_1$ and $\ell_2$ denote mean absolute error and mean squared error, respectively; $\mathcal{L}_{\text{LPIPS}}$ [30] is the perceptual loss; and $\mathcal{L}_{\text{adv}}$ is the adversarial loss with an adaptive weight $w_{\text{adv}}$ that balances reconstruction and adversarial gradients on the decoder's last layer, following VQGAN [7].

For the KL regularization, patch tokens $\boldsymbol{h} \in \mathbb{R}^{N\times d}$ are linearly projected to a mean $\boldsymbol{\mu} \in \mathbb{R}^{N\times d_z}$ and a log-variance $\log \boldsymbol{\sigma}^2 \in \mathbb{R}^{N\times d_z}$, where $d_z = 48$. The latent code is sampled via reparameterization as $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The KL term penalizes the deviation of the diagonal Gaussian posterior from the standard normal prior:

$$\mathcal{L}_{\text{KL}} = \frac{1}{2}\sum_{n,k}\left(\boldsymbol{\mu}_{n,k}^2 + \boldsymbol{\sigma}_{n,k}^2 - 1 - \log \boldsymbol{\sigma}_{n,k}^2\right), \tag{7}$$

weighted by $\lambda_{\text{KL}} = 10^{-6}$ following Rombach *et al.* [29]. We set $\lambda_{\text{MERA}} = 1$ and $\lambda_{\text{SliDR}} = 0.1$ by default; ablation studies on these coefficients can be found in Sec. C.3.

**Two-stage training.** In Stage 1, MedVTok is trained exclusively on 2D images with $\mathcal{L}_{\text{R}}$, $\mathcal{L}_{\text{KL}}$, and $\mathcal{L}_{\text{MERA}}$ ($\lambda_{\text{SliDR}} = 0$). Only 2D expert projectors are active. A diversity sampler ensures that each mini-batch contains at least two samples from every available modality group, with the remaining slots filled from the global pool.

In Stage 2, we initialize from the Stage 1 checkpoint, introduce 3D expert projectors (indices 9–10), and activate $\mathcal{L}_{\text{SliDR}}$ with $\lambda_{\text{SliDR}} = 0.1$. Training alternates between 2D and 3D batches at every

iteration: even iterations process a 2D batch (effective batch size 600) and odd iterations process a 3D batch (effective batch size 48). The 3D dataloader cycles indefinitely to maintain continuous 3D supervision regardless of the smaller 3D dataset size. Within each 3D batch, a balanced sampler maintains an approximately equal ratio of CT and MR volumes.

**Optimization.** We use AdamW [31] for both the tokenizer and discriminator. The tokenizer uses $(\beta_1, \beta_2) = (0.9, 0.95)$ with a cosine annealing schedule: in Stage 1, the learning rate decays from $10^{-4}$ to $10^{-5}$ over $\sim$100K steps; in Stage 2, from $2 \times 10^{-5}$ to $2 \times 10^{-6}$ over $\sim$100K steps. The discriminator uses $(\beta_1, \beta_2) = (0.5, 0.9)$. All training uses BFloat16 mixed precision. During training, images are of mixed resolutions: 256, 384, and 512.

**Compute resources.** Full-scale data training is conducted on 16$\times$ NVIDIA H200 GPUs (141 GB); peak GPU memory usage is 132.54 GB per device. Each compute node is an x86_64 machine with dual Intel Xeon Platinum 8558 processors (2 sockets, 48 physical cores per socket, 2 threads per core). Stage 1 takes approximately 15.75 hours and Stage 2 takes approximately 43.5 hours. The full research project required more compute than reported here due to preliminary and failed experiments during hyperparameter exploration and architecture design.

**License.** The code and model weights will be released under the Apache 2.0 license.

# C Evaluation

## C.1 Evaluation Datasets

Table 8 summarizes all evaluation datasets used in our experiments, grouped by downstream task. We verified that all evaluation datasets listed are excluded from the training corpus.

## C.2 Evaluation Setup and Results

**Token fidelity.** We use autoencoding reconstruction as a proxy task to evaluate token fidelity: each input is encoded and decoded by the frozen tokenizer, and we measure how faithfully the tokens preserve the original signal. For 2D images, evaluation is performed at $512\times512$ resolution. We report PSNR and SSIM between the input and reconstructed images. For 3D volumes, we evaluate on CT and MRI using two datasets. CT intensities are clipped to $[-1000, 1000]$ HU; MRI intensities are clipped at the 0.5th and 99.5th percentiles per volume. PSNR and SSIM are computed at the volume level. Only models with native 3D support (PUMIT, MedVAE, AToken, MedVTok) are included in the 3D evaluation.

Additional visualizations of reconstructions on 2D and 3D modalities are presented in Fig. 6 and Fig. 7, respectively.

**Classification.** We adopt a $k$-nearest neighbor ($k$-NN) protocol on frozen features extracted from the CLS token (or equivalent global representation) of each tokenizer. For tokenizers that do not output such a representation (*e.g.*, SDVAE), we concatenate all features along the channel dimension. The feature vectors are $L_2$-normalized, and classification is performed via softmax-weighted voting over the $k$ nearest neighbors using cosine similarity with a temperature of $T = 0.07$. We sweep $k \in \{10, 20, 100, 200\}$ on the validation set and report the best accuracy for each model. Image resolution is controlled at $256\times256$ when possible.

**Retrieval.** We evaluate image-to-image retrieval using the same frozen features and benchmarks as classification. The training split of each dataset serves as the gallery, and the test split provides the queries. For each query, we compute the cosine similarity between its $L_2$-normalized feature vector and all gallery features, rank the gallery by descending similarity, and report Recall@1 (the fraction of queries whose top-retrieved image shares the same label). Image resolution is controlled at $256\times256$ when possible, consistent with the classification setup.

**Segmentation.** We evaluate the quality of frozen patch tokens for dense prediction by training a lightweight linear segmentation head on top of each tokenizer's output, following the protocol of Siméoni *et al.* [55]. The segmentation head consists of a batch normalization layer followed by a $1\times1$ convolution. For each tokenizer, we train separate heads under identical settings and select the best-performing one by validation Dice score. Training runs for 50 epochs with AdamW (no weight decay), a batch size of 32, cross-entropy loss, and learning rates searched over $\{10^{-2}, 10^{-3}, 10^{-4}\}$.

Table 8: **Evaluation datasets grouped by downstream task.** Rec: reconstruction, Cls: classification, Ret: retrieval, Seg: segmentation, Syn: image synthesis, VQA: visual question answering, MRG: medical report generation. #Cls denotes the number of target classes. Datasets appearing in multiple task groups are listed separately.

| Dataset | Modality | #Samples | #Cls | Task(s) |
|---|---|---|---|---|
| *2D Perception – Reconstruction / Classification / Retrieval* | | | | |
| ChestCTScan [34] | CT | 1,000 | 4 | Rec, Cls, Ret |
| JSIEC [33] | CFP | 1,000 | 39 | Rec, Cls, Ret |
| HAM10K [35] | Dermoscopy | 10,015 | 7 | Rec, Cls, Ret |
| PolypDiag [37] | Endoscopy | 2,202 | 2 | Rec, Cls, Ret |
| BRISC [36] | MRI | 5,023 | 4 | Rec, Cls, Ret |
| OCTDL [38] | OCT | 3,236 | 7 | Rec, Cls, Ret |
| LungHist700 [39] | Pathology | 691 | 3 | Rec, Cls, Ret |
| BUSBRA [40] | Ultrasound | 1,875 | 4 | Rec, Cls, Ret |
| Kermany [41] | X-ray | 5,856 | 2 | Rec, Cls, Ret |
| *2D Perception – Segmentation* | | | | |
| MaplesDR [42] | CFP | 192 | 3 | Seg |
| MosMedDataPlus [43] | CT | 2,729 | 2 | Seg |
| HAM10K [35] | Dermoscopy | 10,015 | 2 | Seg |
| PolypDB [44] | Endoscopy | 3,843 | 2 | Seg |
| BRISC [36] | MRI | 4,746 | 2 | Seg |
| PanNuke [45] | Pathology | 7,901 | 6 | Seg |
| BUSBRA [40] | Ultrasound | 1,875 | 2 | Seg |
| VinDr-RibCXR [46] | X-ray | 245 | 2 | Seg |
| *3D Volumetric Reconstruction* | | | | |
| AMOS-MM [53] | CT | 100 | – | 3D Rec |
| MRISegmenter [54] | MRI | 240 | – | 3D Rec |
| *Image Synthesis* | | | | |
| GMAI-VL-5.5M [1] | Multi-modality | 70,000 | – | Syn |
| *Visual Question Answering (test benchmarks)* | | | | |
| Fundus-MMBench [48] | CFP | 620 | – | VQA |
| GMAI-MMBench [47] | Multi-modality | 4,550 | – | VQA |
| Kvasir-VQA-x1 [51] | Endoscopy | 15,955 | – | VQA |
| MM-Skin [50] | Dermoscopy | 5,481 | – | VQA |
| PathVQA [49] | Pathology | 6,719 | – | VQA |
| SLAKE [52] | Radiography | 2,094 | – | VQA |
| *Medical Report Generation* | | | | |
| AMOS-MM [53] | CT | 1,687 | – | MRG |

Image resolution is controlled at $256\times256$ when possible. We report the overall Dice score over segmentation classes. Table 9 further reports Hausdorff Distance at the 95th percentile (HD95), where MedVTok still leads in terms of average performance across modalities.

**Comparison with modality experts.** Tables 10 and 11 compare MedVTok with the experts used in MERA under the classification and segmentation protocols above. MedVTok exceeds MedImageInsight on seven of nine classification benchmarks and all eight segmentation benchmarks. Among the four specialized experts, it exceeds PeskaVLP in classification and RetCLIP, DermLIP, and PeskaVLP in segmentation. Conch retains higher scores on pathology.

**3D segmentation.** We evaluate frozen tokenizers on AMOS-MM CT [53] and MRISegmenter MRI [54], both excluded from tokenizer pretraining. For each tokenizer, we train a lightweight 3D convolutional head on volumetric features to predict segmentation masks. Table 12 reports overall and per-organ Dice. MedVTok achieves the highest overall Dice on both datasets.

**Cross-dataset 3D segmentation.** We apply the models trained on MRISegmenter directly to AMOS-MM MRI, without training or tuning on the target dataset. Evaluation covers the ten shared anatomical labels. MedVTok achieves the highest transfer Dice (Table 13).

**Image synthesis.** We evaluate how the quality of a tokenizer's latent space influences text-conditioned medical image synthesis. Each frozen tokenizer is integrated into a latent diffusion framework [29] as the only variable: a UNet denoising backbone following the Stable Diffusion v1.4 architecture

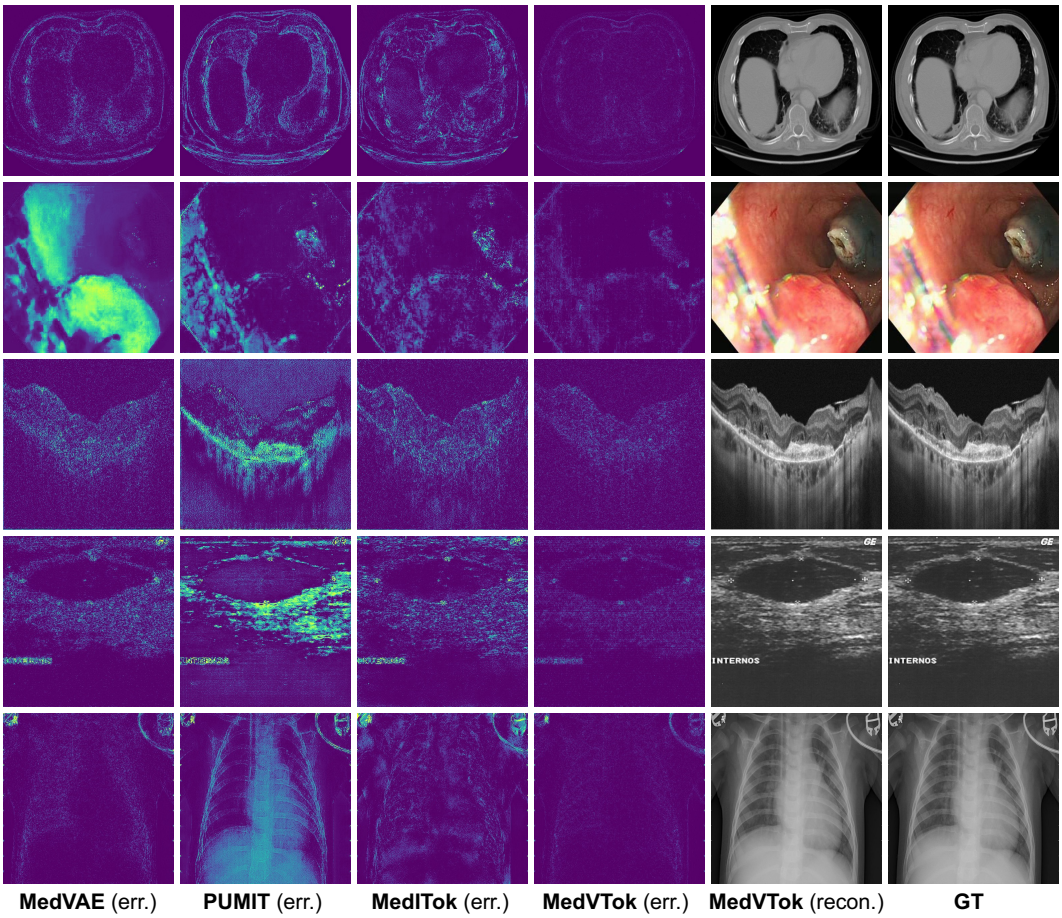



Figure 6: **Additional reconstruction visualization.**

is randomly initialized for each tokenizer (with input/output channels adapted to match the latent dimensionality), and a frozen CLIP-based text encoder provides conditioning. We train separate diffusion models per tokenizer using identical data, architecture, and hyperparameters, ensuring that differences in generation quality are attributable solely to the tokenizer.

Training uses 70,000 images from GMAI-VL-5.5M [1] spanning seven imaging modalities (10,000 images each). The noise scheduler is DDPM with 1,000 timesteps, a scaled-linear beta schedule, and $\epsilon$-prediction parameterization. The model is trained with an effective batch size of 256, the AdamW optimizer, a constant learning rate with 5% linear warmup, and classifier-free guidance dropout of 10%. Note that the purpose of this experiment is to *compare the generative suitability of different tokenizer latent spaces*, rather than to build state-of-the-art text-to-image generative models or to benchmark fine-grained synthesis. The original paired descriptions in GMAI-VL-5.5M are often long and fine-grained; using them directly may introduce confounding factors from text complexity. We therefore simplify the setup, replacing each caption with a controlled template, *e.g.*, "A medical {modality} image," and keep this protocol identical for all tokenizers. Under this setting, differences in generation quality more directly reflect how well each frozen tokenizer organizes medical images into a latent space that is compact, learnable, and amenable to latent diffusion.

At inference, we use DDIM sampling with 100 denoising steps. We generate 500 images per modality and report the Kernel Inception Distance (KID) [64], Fréchet Inception Distance (FID) [56], and Inception Score (IS) [65] for evaluation in Table 14. MedVTok consistently achieves the best average performance across these metrics.

**Visual question answering.** We evaluate each frozen tokenizer as a visual backbone within the LLaVA-Med [58] framework. Given a medical image, the tokenizer encodes it into patch-level tokens,

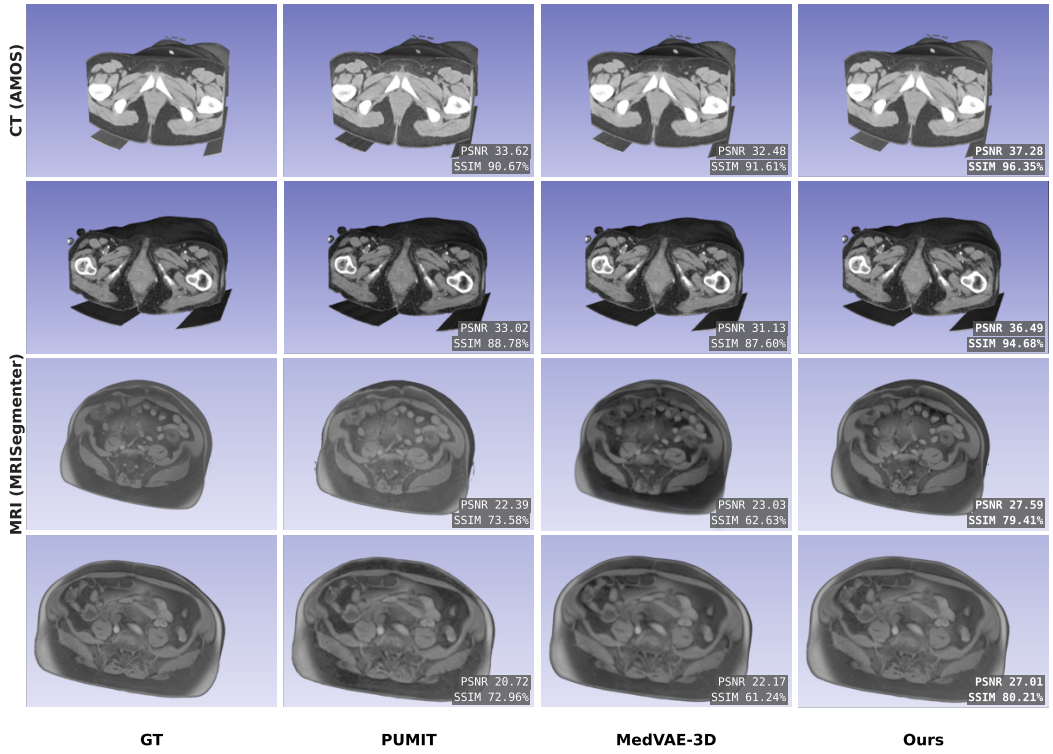



Figure 7: **3D Reconstruction visualization.**

which are then fed through a two-layer MLP connector into Mistral-7B-Instruct as the language model backbone. We train separate LLaVA-Med variants using identical data and hyperparameters while keeping the tokenizer frozen.

Training follows a two-stage protocol. In the first stage, only the MLP connector is trained on approximately 509,369 image–text pairs (learning rate $10^{-3}$, one epoch), with 500,000 pairs from PubMedVision [66] and 9,369 pairs from MM-Skin [50]. In the second stage, LoRA adapters (rank=128, $\alpha$ =256, dropout=0.05) are applied to all linear layers of the language model and fine-tuned jointly with the connector (learning rate $10^{-4}$ for LoRA, $2\times10^{-5}$ for the connector, one epoch), using 140,151 VQA items for training (70,000 from PubMedVision; 21,778 from the MM-Skin training split; 28,719 from the Kvasir-VQA-x1 training split; 19,654 from the PathVQA training split).

At inference, we use sampling with a temperature of 0.2. Model outputs are evaluated via LLM-as-judge binary grading using Qwen3-235B-A22B [67]. We train each LLaVA-Med variant with three random seeds and report accuracy as mean $\pm$ 95% confidence interval in Table 5.

Two representative cases of LLaVA-Med output (integrating MedVTok as the visual encoder) are presented in Fig. 12, where the model identifies the pathological features in the given medical image and answers the question correctly.

**VQA controls.** We hold the connector, language model, data, LoRA settings, decoding, and judge fixed across tokenizers. We also evaluate a frozen language model with only the connector trained, and re-score the standard model outputs with GPT-5. MedVTok leads under connector-only training, and replacing the Qwen3-235B-A22B judge with GPT-5 preserves the tokenizer ranking (Table 15).

**Medical report generation.** We use the AMOS-MM [53] dataset, which provides CT volumes paired with radiology reports covering three anatomical regions (abdomen, chest, and pelvis), with 1,287 pairs for training and 400 pairs for evaluation. CT intensities are windowed to $[-160, 240]$ HU. Each frozen visual tokenizer is integrated into the LaMed framework [59] as the visual backbone: given a 3D volume, the tokenizer first encodes it into latent tokens, which are then fed through a connector (a

Table 9: **Additional results on image segmentation evaluation.** Method type: *semantic encoder; ‡unified tokenizer.

| Metrics | Method | CFP | CT | Derm. | Endo. | MRI | Path. | US | X-ray | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| Dice (↑) | SDVAE [29] | 73.03 | 14.88 | 62.37 | 24.67 | 12.69 | 49.11 | 84.19 | 35.56 | 44.56 |
| | TokenFlow‡ [4] | 70.39 | 19.05 | 72.62 | 40.53 | 30.49 | 52.34 | 91.33 | 49.52 | 53.28 |
| | MingTok‡ [32] | 85.36 | 27.03 | 83.91 | 57.32 | 52.54 | 67.68 | 92.91 | 54.71 | 65.18 |
| | UniTok‡ [10] | 60.93 | 33.13 | 86.25 | 67.25 | 54.90 | 72.51 | 93.35 | 57.60 | 65.74 |
| | AToken‡ [9] | 97.09 | 37.33 | 87.45 | 69.88 | 63.57 | 74.30 | 94.90 | 59.69 | 73.03 |
| | BiomedCLIP* [15] | 50.22 | 31.83 | 87.48 | 60.64 | 57.95 | 73.65 | 95.41 | 59.83 | 64.63 |
| | MedSigLIP* [17] | 97.28 | 40.79 | 87.76 | 69.81 | 67.32 | 76.20 | 95.33 | **62.48** | 74.62 |
| | MedImageInsight* [16] | 92.34 | 25.60 | 85.45 | 66.96 | 55.81 | 42.93 | 95.80 | 50.16 | 64.38 |
| | PUMIT [8] | 40.05 | 25.53 | 77.94 | 37.32 | 36.69 | 66.00 | 90.73 | 54.82 | 53.63 |
| | MedVAE [5] | 61.74 | 12.57 | 57.59 | 19.73 | 10.13 | 36.40 | 81.90 | 30.81 | 38.86 |
| | MedITok‡ [6] | 96.70 | 40.52 | 80.28 | 66.07 | 62.42 | 75.80 | 94.65 | 59.47 | 71.99 |
| | MedVTok‡ | **97.44** | **42.91** | **90.25** | **72.52** | **69.62** | **77.74** | **96.91** | 61.13 | **76.07** |
| HD95 (↓) | SDVAE [29] | 109.69 | 218.98 | 56.30 | 124.25 | 343.33 | 241.99 | 172.77 | 46.39 | 164.21 |
| | TokenFlow‡ [4] | 57.43 | 211.68 | 56.90 | 73.33 | 151.20 | 278.48 | 45.72 | 23.82 | 112.32 |
| | MingTok‡ [32] | 115.80 | 180.54 | 40.70 | 109.14 | 76.67 | 230.27 | 51.49 | 34.18 | 104.85 |
| | UniTok‡ [10] | 104.54 | 123.44 | 18.10 | 51.08 | 66.77 | 162.47 | 22.59 | 13.85 | 70.35 |
| | AToken‡ [9] | 22.99 | 119.80 | 16.05 | 47.32 | 29.93 | 160.72 | 16.64 | **12.63** | 53.26 |
| | BiomedCLIP* [15] | 84.43 | 178.88 | 12.62 | 50.33 | 31.22 | 145.42 | 13.75 | 12.97 | 66.20 |
| | MedSigLIP* [17] | 38.47 | 112.54 | 13.94 | **35.48** | **21.55** | 138.56 | 16.82 | 21.45 | 49.85 |
| | MedImageInsight* [16] | 31.83 | 248.33 | 17.16 | 58.84 | 54.94 | 174.84 | 12.50 | 27.52 | 78.24 |
| | PUMIT [8] | 63.10 | 151.79 | 30.78 | 76.98 | 78.13 | 214.75 | 35.29 | 19.50 | 83.79 |
| | MedVAE [5] | 326.48 | 241.19 | 69.08 | 102.19 | 164.70 | 343.04 | 481.75 | 47.13 | 221.94 |
| | MedITok‡ [6] | 44.53 | 138.61 | 37.06 | 46.60 | 32.97 | 140.30 | 18.31 | 13.85 | 59.03 |
| | MedVTok | **19.12** | **105.06** | **11.67** | 46.31 | 24.76 | **121.30** | **12.43** | 15.21 | **44.48** |

Table 10: **Classification accuracy (%) compared with MERA experts.** Specialized experts are evaluated on their respective modalities. Values are means ± 95% confidence intervals.

| Model | CFP | CT | Derm. | Endo. | MRI | OCT | Path. | US | X-ray |
|---|---|---|---|---|---|---|---|---|---|
| RetCLIP | **91.95**±1.05 | – | – | – | – | – | – | – | – |
| DermLIP | – | – | 79.72±0.66 | – | – | – | – | – | – |
| PeskaVLP | – | – | – | 48.64±2.29 | – | – | – | – | – |
| Conch | – | – | – | – | – | – | **93.08**±2.44 | – | – |
| MedImageInsight | 75.64±2.12 | 75.32±2.83 | **85.87**±0.38 | 44.65±0.74 | **95.47**±0.29 | 90.79±1.18 | 71.79±2.92 | 50.18±2.67 | 96.21±0.83 |
| MedVTok | 90.54±1.05 | **87.03**±1.64 | 75.27±0.99 | **51.00**±1.85 | 95.03±0.52 | **94.00**±1.18 | 78.97±1.65 | **52.92**±1.68 | **97.22**±0.61 |

Table 11: **Segmentation Dice (%) compared with MERA experts.** Specialized experts are evaluated on their respective modalities. Values are means ± 95% confidence intervals.

| Model | CFP | CT | Derm. | Endo. | MRI | Path. | US | X-ray |
|---|---|---|---|---|---|---|---|---|
| RetCLIP | 92.46±1.69 | – | – | – | – | – | – | – |
| DermLIP | – | – | 89.05±1.86 | – | – | – | – | – |
| PeskaVLP | – | – | – | 65.06±1.51 | – | – | – | – |
| Conch | – | – | – | – | – | **83.77**±1.39 | – | – |
| MedImageInsight | 92.34±1.39 | 25.60±1.35 | 85.45±2.10 | 66.96±1.32 | 55.81±1.31 | 42.93±1.86 | 95.80±2.07 | 50.16±1.48 |
| MedVTok | **97.44**±1.59 | **42.91**±1.80 | **90.25**±0.77 | **72.52**±1.15 | **69.62**±1.07 | 77.74±1.56 | **96.91**±1.80 | **61.13**±1.17 |

spatial pooling layer that applies 3D average pooling with 2× downsampling per dimension) and into Llama-3.1-8B-Instruct adapted with LoRA (rank=8, $\alpha$ =16, dropout=0.1).

We train separate LaMed variants, each with a different visual tokenizer as the frozen backbone, using identical data and hyperparameters, and evaluate all variants on the same test set under the same inference settings. Training runs for 15 epochs with AdamW (learning rate $5\times10^{-5}$, cosine decay, 3% warmup), an effective batch size of 4, and BFloat16 precision.

At inference, the model generates region-specific findings autoregressively with a temperature of $T = 0.2$ and top-$p = 0.9$. We also include MedITok [6], a 2D-only tokenizer, for comparison. To adapt it to 3D inputs, we treat the depth dimension as a batch of 2D slices, encode them independently in parallel with the MedITok encoder, and concatenate the resulting tokens along the depth axis before passing them to the connector and LLM.

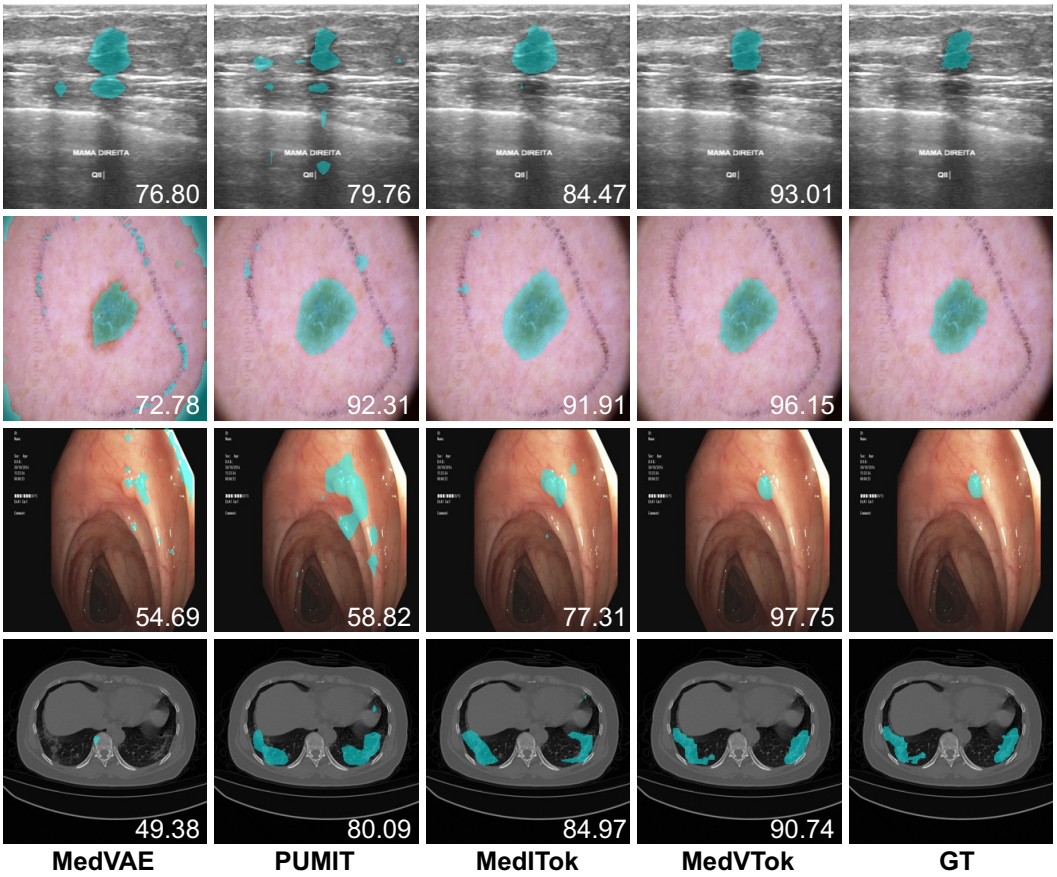



Figure 8: **Qualitative comparison on the image segmentation task.** Segmentation powered by MedVTok achieves the best performance, while other tokenizers lead to false segmentation in the background. In each image, the mask is highlighted in cyan; the Dice score is shown in the lower right corner.

Table 12: **3D segmentation Dice (%).** Each tokenizer is frozen while a lightweight 3D convolutional head is trained. Values include reported 95% confidence intervals.

| | AMOS-MM CT | | | | MRISegmenter MRI | | | |
|---|---|---|---|---|---|---|---|---|
| Organ | PUMIT | MedVAE 3D | AToken | MedVTok | PUMIT | MedVAE 3D | AToken | MedVTok |
| Overall | 28.67 | 29.71 | 36.89 | **50.13** | 44.57±0.76 | 43.70±0.80 | 54.63±0.68 | **57.20**±0.76 |
| Spleen | 63.68±1.58 | 42.94±2.64 | 55.76±2.46 | **70.61**±2.30 | 40.83±2.12 | 40.68±2.14 | 63.28±2.71 | **70.60**±2.85 |
| Kidney | 48.94±1.96 | 45.95±2.36 | 59.56±2.44 | **74.93**±1.74 | 81.60±1.29 | 77.46±1.77 | 83.96±0.67 | **86.48**±0.95 |
| Liver | 76.63±2.68 | 73.73±2.23 | 80.50±1.72 | **86.31**±1.00 | 58.11±1.32 | 59.11±1.23 | 85.71±0.94 | **85.83**±0.70 |
| Pancreas | 13.42±1.99 | 23.51±2.60 | 31.64±2.72 | **41.90**±2.36 | 49.43±2.22 | 43.98±2.08 | 49.23±1.92 | **49.49**±2.71 |
| Adrenal | 16.50±1.30 | 12.74±1.10 | 20.44±1.47 | **23.64**±1.76 | 24.00±1.43 | 21.69±1.87 | **25.63**±1.49 | 23.16±1.64 |
| Gallbladder | 18.78±2.21 | 16.33±2.99 | 14.93±1.82 | **39.69**±4.57 | 24.20±3.46 | 23.31±3.27 | 32.85±3.68 | **34.94**±3.76 |
| Esophagus | 16.29±1.02 | 18.59±2.98 | 15.14±1.80 | **19.04**±2.68 | 48.05±1.82 | 39.28±2.26 | 44.92±1.79 | **52.78**±1.61 |
| Stomach | 4.96±0.76 | 3.71±0.76 | 34.63±3.53 | **62.15**±3.87 | 40.66±2.03 | 46.67±2.05 | 52.08±1.30 | **59.89**±2.30 |
| Duodenum | 3.37±1.73 | 10.44±2.00 | 15.04±1.46 | **30.48**±2.71 | 14.82±1.64 | 16.41±1.62 | **40.11**±1.75 | 39.23±1.80 |
| Major vessels | 24.14±1.82 | 49.19±2.05 | 41.22±1.65 | **52.53**±1.37 | 64.00±1.28 | 68.46±1.18 | 68.52±0.91 | **69.56**±0.93 |

In Tables 6 and 16, we report BLEU-2, ROUGE-L, METEOR, RaTEScore, and GREEN per anatomical region and globally, as mean ± 95% confidence interval over three random seeds. The LaMed model integrating MedVTok as the visual encoder achieves superior performance in terms of the average score on each metric. Representative outputs are shown in Fig. 13, where MedVTok directly encodes the whole CT volume and documents the lesions and abnormalities present.

Table 13: **Cross-dataset 3D segmentation Dice (%).** Models trained on MRISegmenter are evaluated directly on AMOS-MM MRI. Values are means ± 95% confidence intervals.

| Model | MRISegmenter | AMOS-MM MRI |
|---|---|---|
| PUMIT | 44.57±0.76 | 31.44±1.52 |
| MedVAE-3D | 43.70±0.80 | 33.52±3.82 |
| AToken | 54.63±0.68 | 44.39±1.75 |
| MedVTok | **57.20**±0.76 | **47.24**±1.47 |

Table 14: **Additional results on image synthesis evaluation.**

| Metrics | Method | CFP | CT | Derm. | Endo. | MRI | Path. | X-ray | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| KID (↓) | PUMIT [8] | 0.098 | 0.110 | 0.096 | 0.110 | 0.067 | 0.296 | 0.080 | 0.123 |
| | MedVAE [5] | 0.116 | 0.084 | 0.135 | 0.154 | 0.077 | 0.288 | 0.068 | 0.132 |
| | MedITok [6] | 0.256 | 0.099 | 0.078 | **0.024** | 0.054 | **0.055** | 0.055 | 0.089 |
| | MedVTok | **0.027** | **0.076** | **0.077** | 0.035 | **0.048** | 0.139 | **0.019** | **0.060** |
| FID (↓) | PUMIT [8] | 140.40 | 183.49 | 272.26 | 191.49 | 227.64 | 377.12 | 200.14 | 227.51 |
| | MedVAE [5] | 118.13 | 154.42 | 202.09 | 170.29 | 145.25 | 325.87 | 88.55 | 172.08 |
| | MedITok [6] | 90.65 | **142.46** | 176.94 | **58.38** | 122.96 | 99.66 | 77.24 | 109.76 |
| | MedVTok | **59.85** | 146.39 | **142.87** | 82.39 | **116.78** | **94.65** | **69.17** | **101.73** |
| IS (↑) | PUMIT [8] | 1.024 | 1.031 | 1.077 | 1.048 | 1.049 | 1.066 | 1.057 | 1.050 |
| | MedVAE [5] | 1.035 | 1.060 | 1.069 | 1.035 | 1.071 | 1.061 | 1.073 | 1.058 |
| | MedITok [6] | 1.039 | 1.048 | 1.104 | 1.038 | **1.079** | 1.154 | **1.076** | 1.077 |
| | MedVTok | **1.044** | **1.060** | **1.127** | **1.054** | 1.075 | **1.159** | 1.074 | **1.085** |

Table 15: **VQA accuracy (%) under training and judge controls.** Connector-only training freezes the language model and uses Qwen3-235B-A22B for grading. Standard training updates the connector and language-model LoRA adapters. Values are means ± 95% confidence intervals.

| Model | Connector only Qwen3 judge | Standard Qwen3 judge | Standard GPT-5 judge |
|---|---|---|---|
| PUMIT | 23.05±0.44 | 41.94±0.33 | 41.83±0.33 |
| MedVAE | 17.88±0.48 | 39.93±0.41 | 39.58±0.42 |
| MedITok | 24.76±0.39 | 45.39±0.28 | 44.82±0.26 |
| MedVTok | **29.86**±0.36 | **46.48**±0.22 | **46.02**±0.24 |

### C.3 Additional Analyses

**Effect of $\lambda_{\text{MERA}}$.** Fig. 9 shows the effect of the MERA loss weight for values in $\{0, 0.1, 1, 2\}$. Without semantic alignment ($\lambda_{\text{MERA}} = 0$), classification and segmentation scores remain significantly low despite strong reconstruction, as the tokenizer only learns the low-level structure without discriminating the high-level semantics. Increasing $\lambda_{\text{MERA}}$ monotonically improves semantic metrics, but at the cost of reconstruction fidelity: at $\lambda_{\text{MERA}} = 2$, PSNR drops by nearly 3 dB and SSIM falls to 0.868. We set $\lambda_{\text{MERA}} = 1$ as the default, which achieves strong semantic performance while maintaining reconstruction quality.

**Effect of $\lambda_{\text{SliDR}}$.** Fig. 10 shows the effect of the SliDR loss weight. As expected, 2D reconstruction is unaffected by $\lambda_{\text{SliDR}}$ since SliDR operates only on volumetric data. For 3D reconstruction, both PSNR and SSIM improve substantially when SliDR is enabled. Further increasing the weight to $\lambda_{\text{SliDR}} = 0.5$ yields marginal 3D gains but degrades classification accuracy by 2.6%, indicating that excessive volumetric regularization interferes with semantic learning. We therefore adopt $\lambda_{\text{SliDR}} = 0.1$, which captures most of the 3D benefit without compromising other objectives.

**Comparison with alternative depth regularizers.** We compare SliDR with direct gradient matching, total variation (TV), and Laplacian regularization. All variants are initialized from the same 2D-trained MedVTok checkpoint and trained on a controlled subset of 10K volumes. We evaluate volumetric reconstruction on AMOS-MM CT and MRISegmenter MRI (Table 17). SliDR achieves the highest mean PSNR and SSIM on both datasets under this setting.

Table 16: **Additional results on medical report generation evaluation.**

| Metrics | Method | Abdomen | Chest | Pelvis | Avg. |
|---|---|---|---|---|---|
| BLEU (↑) | PUMIT [8] | **0.289**±0.008 | 0.193±0.003 | 0.164±0.007 | 0.215±0.005 |
| | MedVAE-3D [5] | 0.249±0.038 | 0.185±0.020 | 0.152±0.010 | 0.195±0.013 |
| | MedITok [6] | 0.256±0.013 | 0.219±0.020 | **0.294**±0.017 | 0.257±0.012 |
| | MedVTok | 0.272±0.005 | **0.263**±0.045 | 0.288±0.017 | **0.274**±0.009 |
| ROUGE-L (↑) | PUMIT [8] | **0.415**±0.005 | 0.277±0.005 | 0.207±0.007 | 0.300±0.003 |
| | MedVAE-3D [5] | 0.398±0.004 | 0.266±0.003 | 0.183±0.004 | 0.282±0.000 |
| | MedITok [6] | 0.389±0.007 | 0.299±0.025 | 0.407±0.033 | 0.365±0.016 |
| | MedVTok | 0.410±0.006 | **0.353**±0.023 | **0.442**±0.011 | **0.402**±0.005 |
| METEOR (↑) | PUMIT [8] | **0.353**±0.005 | 0.213±0.003 | 0.189±0.007 | 0.252±0.004 |
| | MedVAE-3D [5] | 0.318±0.025 | 0.221±0.010 | 0.172±0.008 | 0.237±0.014 |
| | MedITok [6] | 0.318±0.014 | 0.279±0.004 | 0.388±0.021 | 0.328±0.014 |
| | MedVTok | 0.336±0.005 | **0.327**±0.045 | **0.399**±0.019 | **0.354**±0.007 |
| RaTEScore (↑) | PUMIT [8] | 0.604±0.005 | 0.511±0.035 | 0.201±0.015 | 0.439±0.006 |
| | MedVAE-3D [5] | 0.613±0.022 | 0.497±0.053 | 0.187±0.003 | 0.432±0.011 |
| | MedITok [6] | 0.610±0.006 | 0.501±0.008 | 0.594±0.031 | 0.568±0.015 |
| | MedVTok | **0.619**±0.005 | **0.570**±0.074 | **0.612**±0.006 | **0.600**±0.006 |
| GREEN (↑) | PUMIT [8] | 0.421±0.017 | 0.204±0.010 | 0.124±0.019 | 0.250±0.009 |
| | MedVAE-3D [5] | 0.401±0.024 | 0.186±0.003 | 0.105±0.064 | 0.230±0.179 |
| | MedITok [6] | 0.412±0.025 | 0.256±0.010 | 0.492±0.026 | 0.386±0.013 |
| | MedVTok | **0.425**±0.005 | **0.300**±0.078 | **0.515**±0.044 | **0.413**±0.020 |

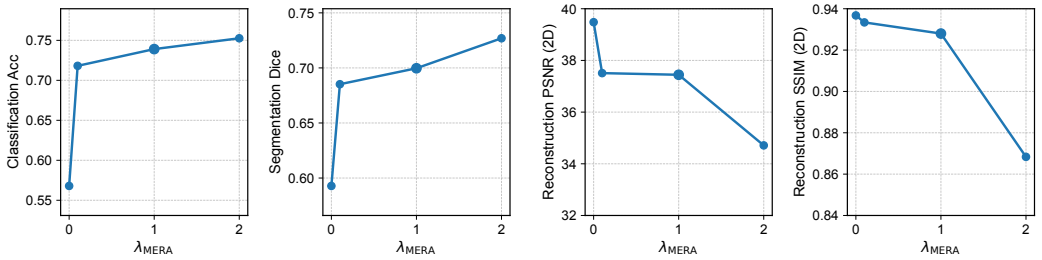



Figure 9: **Ablation study of $\lambda_{\text{MERA}}$.** Classification accuracy and segmentation Dice improve with increasing MERA weight, while reconstruction fidelity (PSNR, SSIM) gradually degrades. $\lambda_{\text{MERA}} = 1$ balances semantic quality and reconstruction.

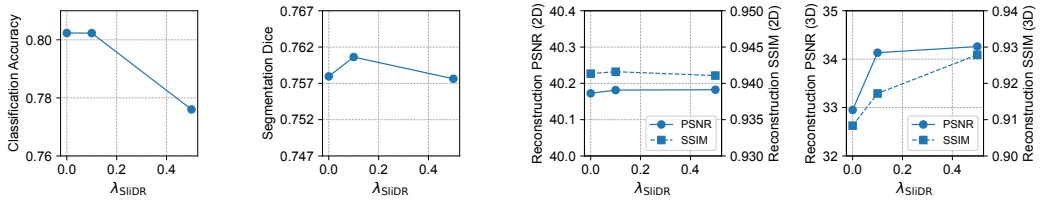



Figure 10: **Ablation study of $\lambda_{\text{SliDR}}$.** SliDR improves 3D reconstruction without affecting 2D fidelity. $\lambda_{\text{SliDR}} = 0.1$ provides the best trade-off, as larger values degrade classification (left).

TV and Laplacian regularization penalize first- and second-order depth differences of the reconstructed volume, respectively, without reference to the ground truth. These penalties encourage smoothness even where the input contains genuine inter-slice changes. SliDR instead matches both orders of slice differentials to the ground-truth volume (Eq. 1), providing a reconstruction target for anatomical transitions.

**Transfer to an unseen modality.** We evaluate frozen tokenizer features on PBC [68], a blood-cell microscopy dataset. This dataset and its hematology modality are absent from tokenizer pretraining.

Table 17: **Comparison of depth regularization objectives.** All variants use the same 2D-trained checkpoint and a controlled 10K-volume training subset. Higher PSNR and SSIM indicate better reconstruction. The best mean in each column is marked in **bold**.

| | AMOS-MM CT | | MRISegmenter MRI | |
|---|---|---|---|---|
| Objective | PSNR (dB) | SSIM (%) | PSNR (dB) | SSIM (%) |
| Direct gradient | 33.24±0.22 | 93.01±0.25 | 26.11±0.14 | 78.62±0.44 |
| TV | 33.05±0.21 | 93.08±0.24 | 26.04±0.14 | 78.00±0.45 |
| Laplacian | 32.99±0.21 | 93.08±0.24 | 26.20±0.15 | 77.86±0.45 |
| SliDR | **33.46**±0.22 | **93.21**±0.25 | **26.23**±0.15 | **78.70**±0.44 |

Table 18: **Classification accuracy (%) on the unseen PBC modality.** Both protocols use frozen tokenizer features. Values are means ± 95% confidence intervals.

| Model | $k$-NN | Linear probe |
|---|---|---|
| MedVAE | 40.47±0.15 | 79.47±1.35 |
| PUMIT | 36.98±0.55 | 83.25±0.84 |
| MedITok | 90.73±0.62 | 94.62±0.76 |
| AToken | 92.00±2.07 | 92.84±0.87 |
| MedVTok | **94.78**±1.81 | **96.84**±0.59 |

Table 19: **Comparison on the same 2M training images.**

| Model | PSNR (dB) | Accuracy (%) | Dice (%) |
|---|---|---|---|
| AToken | 39.35 | 63.65 | 69.57 |
| MedITok | 36.83 | 66.59 | 67.01 |
| MedVTok | **39.71** | **71.70** | **73.21** |

We use $k$-NN classification and train a linear classifier on frozen features. MedVTok achieves the highest accuracy under both protocols: 94.78% and 96.84%, respectively (Table 18).

**Comparison on the same training data.** For controlled comparison with state-of-the-art tokenizers, we retrain AToken, MedITok, and MedVTok on MedITok's 2M image–text dataset using the same training settings. AToken and MedITok use text supervision; MedVTok uses the images alone. MedVTok achieves the highest reconstruction PSNR, classification accuracy, and segmentation Dice in this comparison (Table 19).

**$t$-SNE analysis.** To qualitatively assess the discriminative structure of the learned representations, we visualize $t$-SNE embeddings of frozen tokenizer features in Fig. 11. We compare PUMIT, MedVAE, MedITok, and MedVTok on two datasets: (a) a microscopy blood cell dataset PBC [68] (covering multiple cell types: basophil, eosinophil, erythroblast, immature granulocytes, lymphocyte, monocyte, neutrophil, platelet), which is an out-of-distribution modality not seen during training, and (b) the JSIEC [33] fundus dataset. On the PBC dataset, MedVTok produces clearly separated clusters for each cell type despite never seeing microscopy images during training, indicating strong generalization of the learned semantic structure. On JSIEC, MedVTok forms tighter, more coherent clusters for fine-grained retinal conditions compared to the baselines, which exhibit more inter-class overlap. These visualizations corroborate the quantitative classification and retrieval results, confirming that MedVTok encodes semantically meaningful representations that generalize across both in-distribution and out-of-distribution modalities.

## D Broader Impact

A general-purpose medical visual tokenizer can accelerate multimodal medical AI research and patient-level modeling by providing a shared, reusable visual interface across imaging modalities. This may lower the barrier for building clinical decision-support tools, particularly for under-resourced specialties or institutions that lack the data and compute. By open-sourcing the model and code, we aim to promote reproducibility and equitable access to medical AI infrastructure.

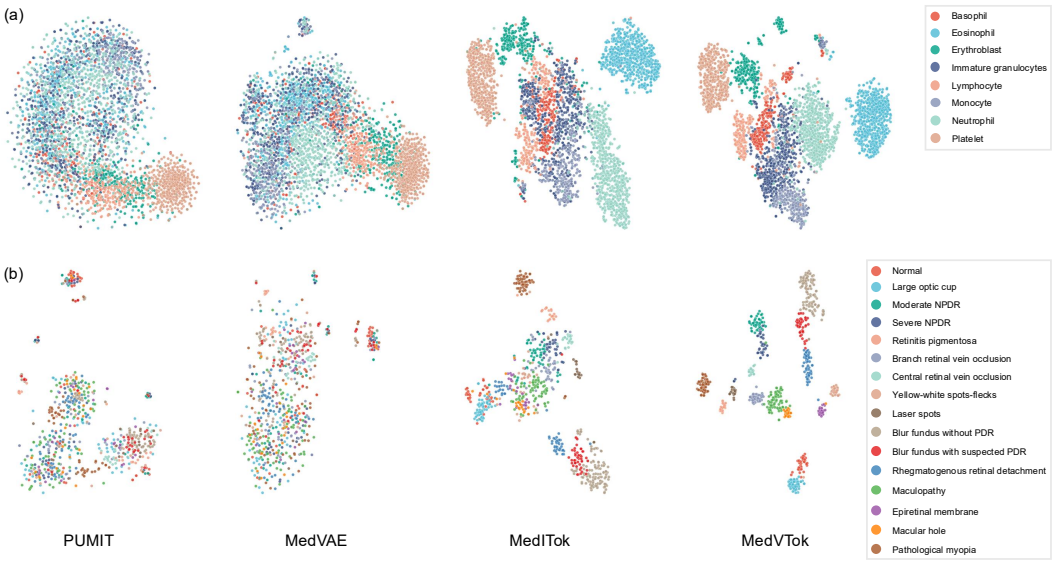



Figure 11: $t$**-SNE visualization of frozen tokenizer representations.** (a) Out-of-distribution microscopy blood cell dataset. (b) JSIEC fundus dataset (showing the top 16 retinal conditions). MedVTok produces the most clearly separated and coherent clusters in both settings.

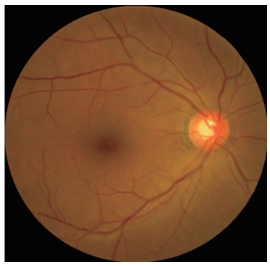
As with any medical AI component, there is a risk that downstream systems built on MedVTok could produce incorrect predictions that, if acted upon without appropriate clinical oversight, may lead to patient harm. We will release MedVTok under the Apache 2.0 license with clear documentation stating that it is intended for research use. We encourage downstream users to rigorously validate performance on their target populations and to comply with applicable regulations before any clinical deployment.



Can you identify any other features in the image that might indicate the presence of melanoma?

Yes, there are several features that could suggest the presence of melanoma. For example, the lesion has irregular borders, which is another common characteristic of melanoma. Additionally, the lesion appears to be asymmetrical, another hallmark of melanoma.

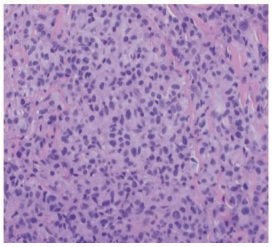
MedVTok



What is the grading of diabetic retinopathy in the fundus color photo?
A: proliferative diabetic retinopathy.
B: no diabetic retinopathy.
C: moderate nonproliferative diabetic retinopathy.
D: mild nonproliferative diabetic retinopathy.
E: severe nonproliferative diabetic retinopathy.

D: Mild nonproliferative diabetic retinopathy.

MedVTok

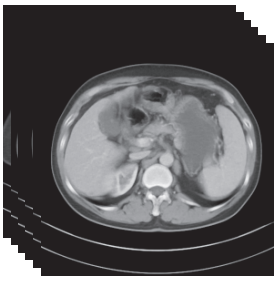


Are there any structures in the image that suggest inflammation or infection?

Yes, there are structures in the image that suggest inflammation or infection. For example, there is a dense cellular infiltrate with some small dark cells that may represent lymphocytes. This indicates that the body's immune system is responding to some kind of threat, which can be seen in inflammation.

MedVTok

Figure 12: Representative responses from the LLaVA-Med model integrating MedVTok as the visual backbone.

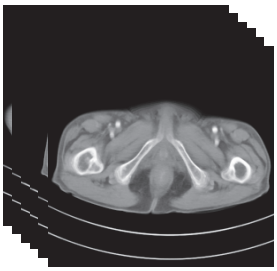


**Pelvis**: The bladder is filled, with no thickening of the wall. The prostate is not enlarged. No ascites is observed. The uterus is normal, with normal density and no abnormalities. The intestine is normal. The bladder-vesical junction is clear. The pelvic region is normal, with no soft tissue mass and no lymph nodes seen in the pelvic cavity.

**Abdomen**: A low-density lesion adjacent to the diaphragm is seen in the right posterior lobe of the liver. Within the right kidney, a punctate high-density lesion is noted in the lower calyx along with a small lesion at the posterior margin. There is no dilation of the intrahepatic or extrahepatic bile ducts. The gallbladder is not enlarged, with no thickening of the wall and no abnormal density inside. The spleen, pancreas, and pancreatic duct show normal morphology, size, and density. No signs of ascites are observed.

**Chest**: The lung fields are clear and normal with no evidence of consolidation. No obvious abnormalities in the seminal vesicles. Chest bones are normal. No pleural effusion is seen in both pleural cavities or bilateral pleural cavities. No enlargement is observed in the bilateral pulmonary hila. No thickening of the pleura is observed. The shape and size of the heart and great vessels are normal.

MedVTok



**Pelvis**: The bladder is inadequately filled, with no definite abnormal density in the cavity and no apparent bladder wall thickening. A punctate hyperdense focus is noted in the prostate. Multiple small retroperitoneal lymph nodes are present, without definite abnormal enhancement on plain or contrast-enhanced imaging.

**Abdomen**: A soft-tissue mass is present in the pancreatic region with indistinct margins from the pancreas and adjacent bowel, associated with irregular bowel wall thickening. There is dilation with fluid in the intrahepatic and extrahepatic bile ducts and the pancreatic duct. The pancreatic body and tail show no enhancement. Multiple round low-density lesions are seen in both kidneys, up to about 19 mm, with clear margins; some show slightly higher density and peripheral punctate calcification. The liver is normal in size and shape with homogeneous density and no focal enhancing lesions.

**Chest**: Cardiac size and contour are normal. No enlarged lymph nodes are identified. There is focal calcified thickening of the right pleura and a small left pleural effusion. Opacities with clear margins are seen in both lungs; the remaining lung fields are clear. The trachea and major bronchi are patent. The aorta and coronary arteries show linear calcifications. No pericardial abnormality is seen.

MedVTok

Figure 13: Representative medical reports generated by the LaMed model integrating MedVTok as the visual backbone.

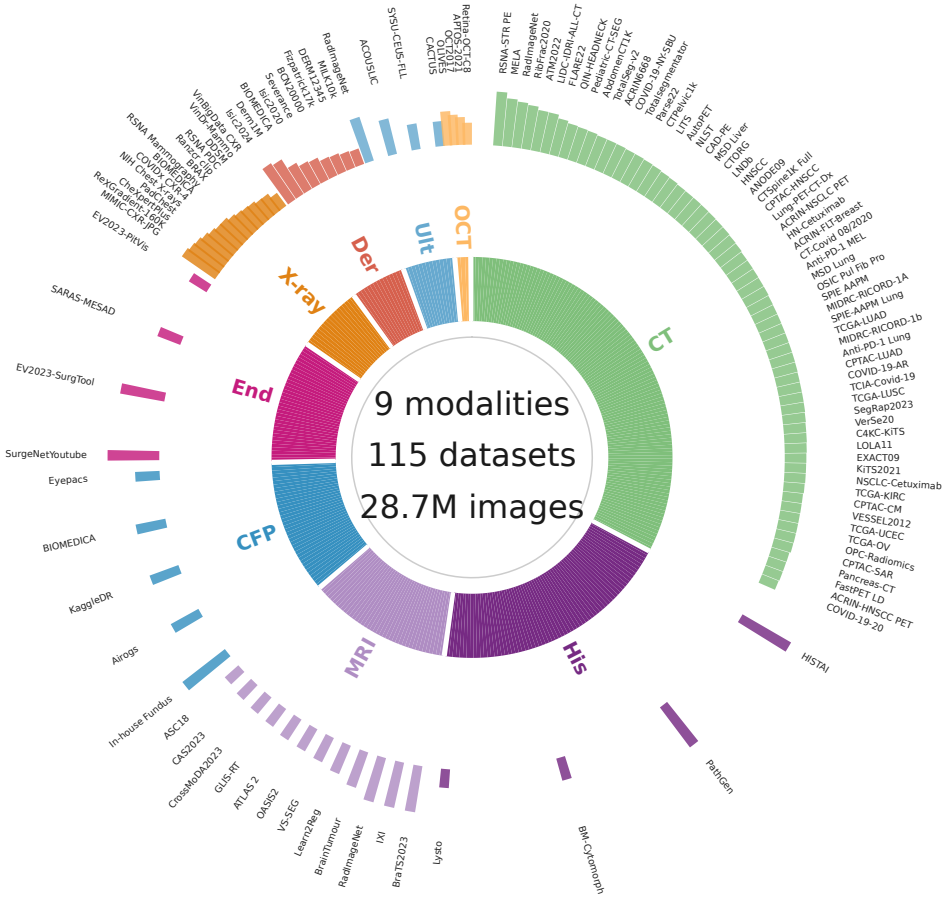



Figure 14: **Training data distribution across imaging modalities (2D corpus).** 2D corpus comprising 28.7M images across 9 modalities and 115 datasets. The colored mid-ring encodes modality with arc length proportional to its sample count; outer spokes correspond to source datasets, with spoke length scaled by $\log_{10}(\text{count})$. Cross-modality datasets (*e.g.*, RadImageNet, BIOMEDICA, GLIS-RT) are split equally across their constituent modalities. Long dataset names are abbreviated for layout; full names are listed in Table 20.

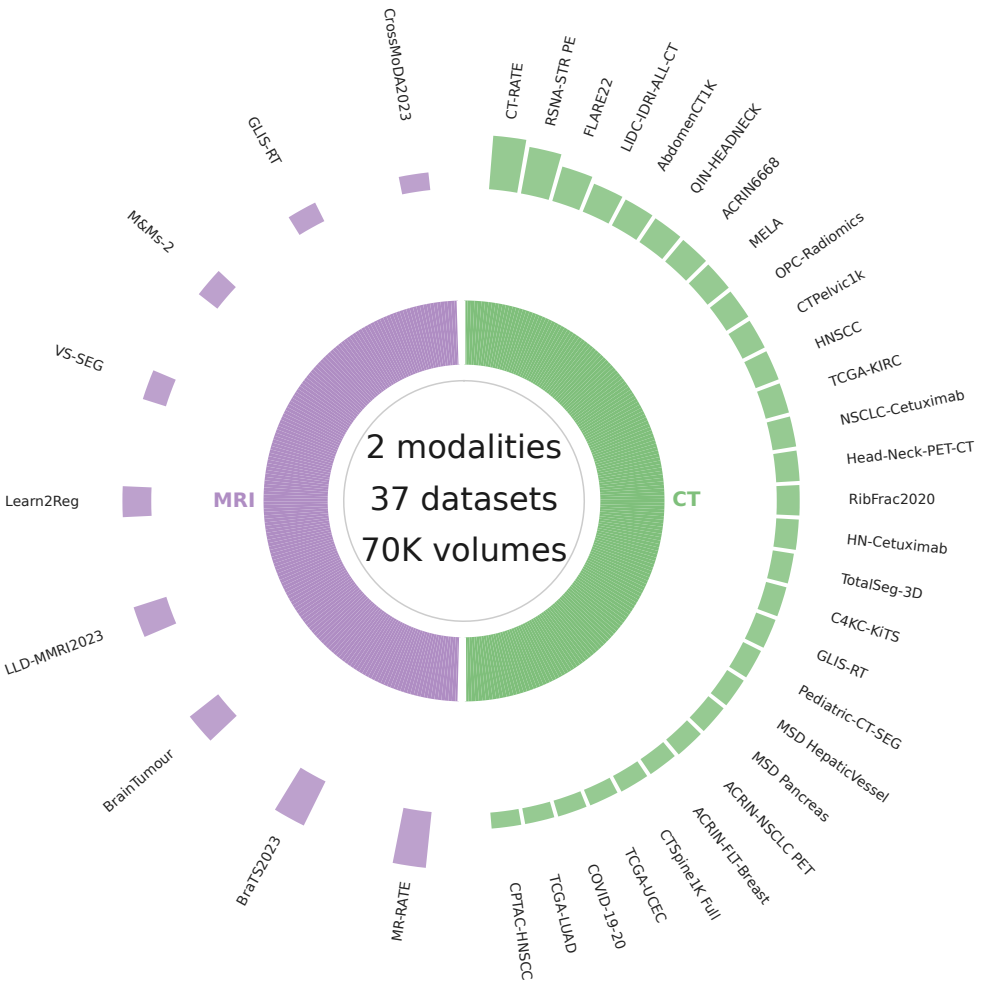



Figure 15: **Training data distribution across imaging modalities (3D corpus).** 3D corpus comprising 70K volumes across 2 modalities and 37 datasets. The colored mid-ring encodes modality with arc length proportional to its sample count; outer spokes correspond to source datasets, with spoke length scaled by $\log_{10}(\text{count})$. Cross-modality datasets are split equally across their constituent modalities. Long dataset names are abbreviated for layout; full names are listed in Table 21.

Table 20: 2D training datasets.

| Dataset | Modality | Count |
|---|---|---|
| RSNA STR Pulmonary Embolism Detection [69] | CT | 2,854,455 |
| MELA [70] | CT | 1,198,641 |
| RibFrac2020 [71] | CT | 675,992 |
| ATM2022 [72] | CT | 419,489 |
| LIDC-IDRI-ALL-CT [73] | CT | 403,462 |
| FLARE22 [74] | CT | 236,577 |
| QIN-HEADNECK [75] | CT | 171,846 |
| Pediatric-CT-SEG [76] | CT | 156,111 |
| AbdomenCT1K [77] | CT | 145,254 |
| TotalSegmentator v2 [78] | CT | 123,420 |
| ACRIN6668 [79] | CT | 121,154 |
| COVID-19-NY-SBU [80] | CT | 116,764 |
| Totalsegmentator [78] | CT | 107,256 |
| Parse22 [81] | CT | 104,622 |
| CTPelvic1k [82] | CT | 96,949 |
| LITS [83] | CT | 90,524 |
| AutoPET [84] | CT | 81,135 |
| NLST [85] | CT | 76,000 |
| CAD-PE [86] | CT | 66,956 |
| MSD Liver [87] | CT | 65,104 |
| CTORG [88] | CT | 64,904 |
| LNDb [89] | CT | 62,969 |
| HNSCC [90] | CT | 61,077 |
| ANODE09 [91] | CT | 58,647 |
| CTSpine1K Full [92] | CT | 57,145 |
| CPTAC-HNSCC [93] | CT | 52,300 |
| Lung-PET-CT-Dx [94] | CT | 49,055 |
| ACRIN-NSCLC-FDG-PET [95] | CT | 42,772 |
| Head-Neck Cetuximab [96] | CT | 42,294 |
| ACRIN-FLT-Breast [97] | CT | 41,063 |
| CT-Covid-19-August2020 [98] | CT | 39,901 |
| Anti-PD-1 MELANOMA [99] | CT | 39,852 |
| MSD Lung [87] | CT | 36,684 |
| OSIC Pul Fib Pro [100] | CT | 35,231 |
| SPIE AAPM [101] | CT | 34,776 |
| MIDRC-RICORD-1A [102] | CT | 33,928 |
| SPIE-AAPM-Lung-CT-Challenge [101] | CT | 33,578 |
| TCGA-LUAD [103] | CT | 33,221 |
| MIDRC-RICORD-1b [102] | CT | 32,953 |
| Anti-PD-1 Lung [104] | CT | 32,388 |
| CPTAC-LUAD [105] | CT | 30,882 |
| COVID-19-AR [106] | CT | 30,304 |
| TCIA-Covid-19 [107] | CT | 30,304 |
| TCGA-LUSC [108] | CT | 29,954 |
| SegRap2023 [109] | CT | 27,932 |
| VerSe20 [110] | CT | 25,860 |
| C4KC-KiTS [111] | CT | 25,598 |
| LOLA11 [112] | CT | 24,156 |
| EXACT09 [113] | CT | 22,747 |
| KiTS2021 [114] | CT | 22,176 |
| NSCLC-Cetuximab [115] | CT | 21,911 |
| TCGA-KIRC [116] | CT | 20,630 |
| CPTAC-CM [117] | CT | 18,513 |
| VESSEL2012 [118] | CT | 18,289 |
| TCGA-UCEC [119] | CT | 17,924 |
| TCGA-OV [120] | CT | 17,917 |
| OPC-Radiomics [121] | CT | 17,853 |
| CPTAC-SAR [122] | CT | 17,502 |
| Pancreas-CT [123] | CT | 14,485 |
| FastPET LD [124] | CT | 14,047 |
| ACRIN-HNSCC-FDG-PET-CT [125] | CT | 13,989 |
| COVID-19-20 [126] | CT | 12,827 |
| RadImageNet [127] | CT, MRI, Ultrasound | 2,639,575 |

*Continued on next page*

| Dataset | Modality | Count |
|---|---|---|
| Isic2024 [128] | Dermoscopy | 618,536 |
| Derm1M [21] | Dermoscopy | 412,886 |
| Isic2020 [129] | Dermoscopy | 44,108 |
| Severance [130] | Dermoscopy | 35,170 |
| BCN20000 [131] | Dermoscopy | 18,946 |
| Fitzpatrick17k [132] | Dermoscopy | 16,577 |
| DERM12345 [133] | Dermoscopy | 12,345 |
| MILK10k [134] | Dermoscopy | 10,480 |
| SurgeNetYoutube [135] | Endoscopy | 2,061,065 |
| EndoVis2023-Surgtoolloc [136] | Endoscopy | 745,751 |
| SARAS-MESAD [137] | Endoscopy | 35,132 |
| EndoVis2023-PitVis [138] | Endoscopy | 19,967 |
| In-house Fundus Dataset | CFP | 2,811,937 |
| Airogs [139] | CFP | 101,442 |
| KaggleDR [140] | CFP | 88,712 |
| Eyepacs [141] | CFP | 35,126 |
| BIOMEDICA [142] | CFP, Dermoscopy, X-ray | 253,443 |
| HISTAI [143] | Histopathology | 4,000,000 |
| PathGen [144] | Histopathology | 1,606,060 |
| Bone Marrow Cytomorphology [145] | Histopathology | 27,783 |
| Lysto [146] | Histopathology | 14,920 |
| BraTS2023 [147] | MRI | 973,066 |
| IXI [148] | MRI | 924,426 |
| BrainTumour [149] | MRI | 246,671 |
| Learn2Reg [150] | MRI | 89,572 |
| Vestibular-Schwannoma-SEG [151] | MRI | 51,574 |
| OASIS2 [152] | MRI | 39,741 |
| ATLAS 2 [153] | MRI | 36,029 |
| GLIS-RT [154] | MRI | 24,979 |
| CrossMoDA2023 [155] | MRI | 21,981 |
| CAS2023 [156] | MRI | 19,200 |
| ASC18 [157] | MRI | 13,552 |
| OLIVES [158] | OCT | 152,346 |
| OCT2017 [159] | OCT | 84,484 |
| APTOS-2021 [160] | OCT | 56,972 |
| Retina-OCT-C8 [161] | OCT | 24,000 |
| ACOUSLIC [162] | Ultrasound | 210,929 |
| SYSU-CEUS-FLL-intv1 [163] | Ultrasound | 43,611 |
| CACTUS [164] | Ultrasound | 37,736 |
| RSNA Mammography [165] | X-ray | 54,710 |
| DDSM [166] | X-ray | 20,800 |
| VinDr-Mammo [167] | X-ray | 19,982 |
| MIMIC-CXR-JPG [168] | X-ray | 372,693 |
| ReXGradient-160K [169] | X-ray | 273,004 |
| CheXpertPlus [170] | X-ray | 223,315 |
| PadChest [171] | X-ray | 160,839 |
| NIH Chest X-rays [172] | X-ray | 112,120 |
| COVIDx CXR-4 [173] | X-ray | 84,817 |
| BRAX [174] | X-ray | 40,944 |
| Ranzcr clip [175] | X-ray | 33,665 |
| RSNA PDC [176] | X-ray | 29,684 |
| VinBigData CXR [177] | X-ray | 9,850 |
| **Total** | | **28,769,502** |

Table 21: 3D training datasets.

| Dataset | Modality | Count |
|---|---|---|
| CT-RATE [178] | CT | 15,000 |
| RSNA STR Pulmonary Embolism Detection [69] | CT | 7,148 |
| FLARE22 [74] | CT | 1,874 |
| LIDC-IDRI-ALL-CT [73] | CT | 967 |
| AbdomenCT1K [77] | CT | 882 |
| QIN-HEADNECK [75] | CT | 816 |
| ACRIN6668 [79] | CT | 764 |
| MELA [70] | CT | 688 |
| OPC-Radiomics [121] | CT | 606 |
| CTPelvic1k [82] | CT | 525 |
| HNSCC [90] | CT | 507 |
| TCGA-KIRC [116] | CT | 474 |
| NSCLC-Cetuximab [115] | CT | 465 |
| Head-Neck-PET-CT [179] | CT | 443 |
| RibFrac2020 [71] | CT | 417 |
| Head-Neck Cetuximab [96] | CT | 399 |
| Totalsegmentator dataset [78] | CT | 346 |
| C4KC-KiTS [111] | CT | 335 |
| Pediatric-CT-SEG [76] | CT | 296 |
| MSD HepaticVessel [87] | CT | 286 |
| MSD Pancreas [87] | CT | 280 |
| ACRIN-NSCLC-FDG-PET [95] | CT | 278 |
| ACRIN-FLT-Breast [97] | CT | 238 |
| CTSpine1K Full [92] | CT | 229 |
| TCGA-UCEC [119] | CT | 205 |
| COVID-19-20 [126] | CT | 199 |
| TCGA-LUAD [103] | CT | 188 |
| CPTAC-HNSCC [93] | CT | 186 |
| GLIS-RT [154] | CT, MRI | 644 |
| MR-RATE [180] | MRI | 20,000 |
| BraTS2023 [147] | MRI | 8,956 |
| BrainTumour [149] | MRI | 1,936 |
| LLD-MMRI2023 [181] | MRI | 1,580 |
| Learn2Reg [150] | MRI | 828 |
| Vestibular-Schwannoma-SEG [151] | MRI | 484 |
| M&Ms-2 [182] | MRI | 448 |
| CrossMoDA2023 [155] | MRI | 227 |
| **Total** | | **70,144** |

## 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: Contributions are stated in the abstract and Sec. 1, and supported by experiments in Sec. 4.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: Limitations are discussed in Sec. 5.

   Guidelines:

   - The answer [N/A] 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: [N/A]

Justification: The paper does not include formal theorems or proofs; Sec. 3 presents loss formulations only.

Guidelines:

- The answer [N/A] 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: Architecture details, hyperparameters, training schedule, and loss weights are provided in Sec. 4 and the Appendix.

   Guidelines:

   - The answer [N/A] 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: Preview code is at the supplementary material. Full code and model weights will be released publicly upon acceptance. Implementation details are described in detail in the main text and appendix.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.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://neurips.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) necessary to understand the results?

Answer: [Yes]

Justification: Optimizer, learning rates, batch sizes, loss weights, and training schedule are specified in Sec. 4 and the Appendix.

Guidelines:

- The answer [N/A] 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: [Yes]

Justification: 95% confidence intervals over 3 random seeds are reported for VQA, and report generation experiments.

Guidelines:

- The answer [N/A] 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: GPU type, count, and training time are reported in the Appendix.

   Guidelines:

   - The answer [N/A] 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: The research uses publicly available medical imaging datasets and complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] 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: A broader impact discussion is included in the Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The medical visual tokenizer does not pose a high risk for misuse.

    Guidelines:

    - The answer [N/A] 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: All datasets and pretrained models are cited with their original sources throughout the paper.

    Guidelines:

    - The answer [N/A] 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` 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: [Yes]

    Justification: The training corpus composition and model architecture are documented; code and weights will be released with a license upon acceptance.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This work does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This work does not involve human subjects research.

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are not part of the core tokenizer method but are used in downstream evaluation: Mistral-7B-Instruct as the LLaVA-Med backbone for VQA, Llama-3.1-8B-Instruct for report generation, and Qwen3-235B-A22B as an LLM-as-Judge for VQA accuracy grading. Their usage is described in Sec. 4 and the Appendix.

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.