

# MEDITok: A UNIFIED TOKENIZER FOR MEDICAL IMAGE SYNTHESIS AND INTERPRETATION

Anonymous authors

Paper under double-blind review

## ABSTRACT

Advanced autoregressive models have reshaped multimodal AI. However, their transformative potential in medical imaging remains largely untapped due to the absence of a *unified* visual tokenizer—one capable of capturing fine-grained visual structures for faithful image reconstruction and realistic image synthesis, as well as rich semantics for accurate diagnosis and image interpretation. To this end, we present MedITok, the first unified tokenizer tailored for medical images, encoding both low-level structural details and high-level clinical semantics within a unified latent space. To balance these competing objectives, we introduce a novel two-stage training framework: a visual representation alignment stage that cold-starts the tokenizer reconstruction learning with a visual semantic constraint, followed by a textual semantic representation alignment stage that infuses detailed clinical semantics into the latent space. Trained on the meticulously collected large-scale dataset with over 30 million medical images and 2 million image-caption pairs, MedITok achieves state-of-the-art performance on more than 30 datasets across 9 imaging modalities and 4 different tasks. By providing a unified token space for autoregressive modeling, MedITok supports a wide range of tasks in clinical diagnostics and generative healthcare applications. Model and code are available in the supplementary material.

## 1 INTRODUCTION

The rapid evolution of advanced autoregressive (AR) models, such as GPT-4o (OpenAI, 2025), has revolutionized multimodal learning. These models excel at generating and understanding text, image, and audio data via unified processing of token-based representations. In medical imaging, AR models begin to demonstrate similar promise, powering report generation (Tanno et al., 2025), tumor segmentation (Chen et al., 2025a), counterfactual synthesis (Ma et al., 2025a), and diagnostic visual question answering (VQA) (Li et al., 2023). By translating complex biomedical image patterns into token sequences, these models can synthesize realistic images and interpret clinical cues (e.g., ground-glass opacities on chest computed tomography, microcalcifications on mammography) in the images, with the potential to streamline workflows and improve patient outcomes.

A critical ingredient in building a powerful AR model is the *visual tokenizer*, which translates an input image to a sequence of discrete tokens suitable for AR modeling. Existing approaches can be divided into two categories. (1) Generation-oriented tokenizers optimized for pixel-level reconstruction<sup>1</sup>, e.g., VQGAN (Esser et al., 2021). These tokenizers precisely capture low-level structure in the image that is vital to image compression (Varma et al., 2025; Wang et al., 2024c) and

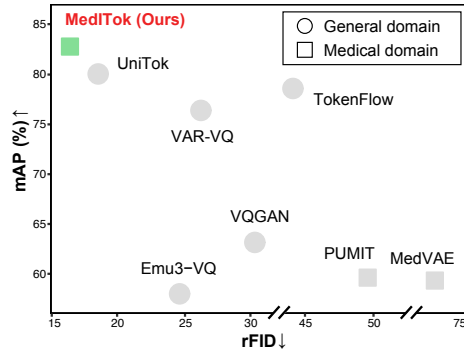


Figure 1: Performance comparison of different tokenizers on medical image reconstruction (rFID) and classification (mAP). MedITok achieves the best of both worlds.

<sup>1</sup>In this paper, “reconstruction” refers to autoencoding reconstruction: decoding an input image from its latent representation.

generation (Zhu et al., 2024; Sun et al., 2024; Yu et al., 2024; Yao et al., 2025). However, they do not explicitly encode discriminative features and are therefore not suitable for interpreting the concepts and objects embedded in the image. (2) Interpretation-driven tokenizers trained with discriminative objectives, *e.g.*, CLIP (Radford et al., 2021). These tokenizers excel at capturing high-level textual semantics, making them indispensable for visual comprehension, but they fail to accurately retain spatial structures and textures in the image.

**Motivation.** Visual tokens that embed only one side of this structure-semantic spectrum will offload the representation learning burden onto downstream AR models, which often incurs heavy pre-training costs and can still leave either side under-utilized (Wang et al., 2024b; Chen et al., 2025b). These limitations are especially acute in the medical domain, where clinical tasks typically demand both precise visual structures and clinical semantics. However, current medical image tokenizers tend to specialize in one single aspect (Luo et al., 2023b; Zhang et al., 2023b), which lacks a unified, information-rich token space and thereby limits the potential of downstream medical AR models for accurate, interpretable, and data-efficient diagnosis.

Our goal is to democratize a foundation visual tokenizer for medical images. Nonetheless, training a unified tokenizer for medical images poses unique challenges: (1) A naïve joint optimization of visual reconstruction and textual semantic objectives often causes mutual interference and degraded performance (Wu et al., 2025; Qu et al., 2024). (2) Paired image-caption data for training is much more scarce in the medical domain, compared to the abundant unlabeled images.

To address these issues, we propose a novel two-stage training framework. Instead of directly coupling the visual reconstruction and textual semantic, it involves a *visual representation alignment stage* to first establish basic semantic awareness with strong reconstruction capabilities as a cold-start, followed by the *textual semantic alignment stage* for learning finer-grained semantic information. This framework leads to our model: MedITok, the first unified visual tokenizer tailored for medical images. MedITok encodes both low-level structural information, supporting image synthesis and compression, and high-level semantics, enabling medical image interpretation and multimodal comprehension, serving as a general foundation for diverse community use.

Specifically, the first training stage cold-starts MedITok on pure medical images, optimizing for reconstruction fidelity with a light semantic constraint on the latent space. Then, the textual semantic alignment stage tunes MedITok on image-caption pairs, enhancing semantic richness by aligning visual tokens to textual embeddings of detailed captions. This approach allows MedITok to *effectively encode structural and semantic information while exploiting both unpaired medical images and image-text pairs at scale*, making a unified token space to develop powerful AR models for diverse tasks. To achieve this, we meticulously collect a large-scale dataset comprising over 30 million medical images and 2 million image-caption pairs from more than 300 public sources, with broad coverage of imaging modalities, anatomies, and pathologies. This collection ensures that MedITok learns robust representations for medical image synthesis and interpretation.

**Contributions.** (1) We propose a novel training framework for developing a unified visual tokenizer, which effectively scales up with medical image and text data and progressively builds a unified token space. (2) We introduce MedITok, the first medical image tokenizer that unifies the encoding of structural details and clinical semantics. (3) Extensive experimental results on over 30 datasets, spanning 9 imaging modalities, across 4 different tasks, showcase MedITok’s state-of-the-art performance over existing approaches and broad applicability to diverse medical tasks. (4) Model and code will be open-source. Data access links are provided respecting all original licenses.

## 2 RELATED WORK

**AR Models in Medical Vision Tasks.** AR models have shown impressive scalability and generalizability in general vision-language processing. In medical domain, these models have been extended to a variety of tasks: image captioning and VQA for interpreting scans and assist diagnosis (Li et al., 2023; Moor et al., 2023; Chen et al., 2024c), lesion segmentation model across different imaging modalities (Chen et al., 2025a), medical image synthesis for counterfactual analysis (Ma et al., 2025a) and modality transfer (Ren et al., 2024), *etc.* More recently, HealthGPT (Lin et al., 2025) further unifies both medical image synthesis and comprehension capabilities within an AR framework for broader applications. However, these methods typically general-domain tokenizers pre-trained on

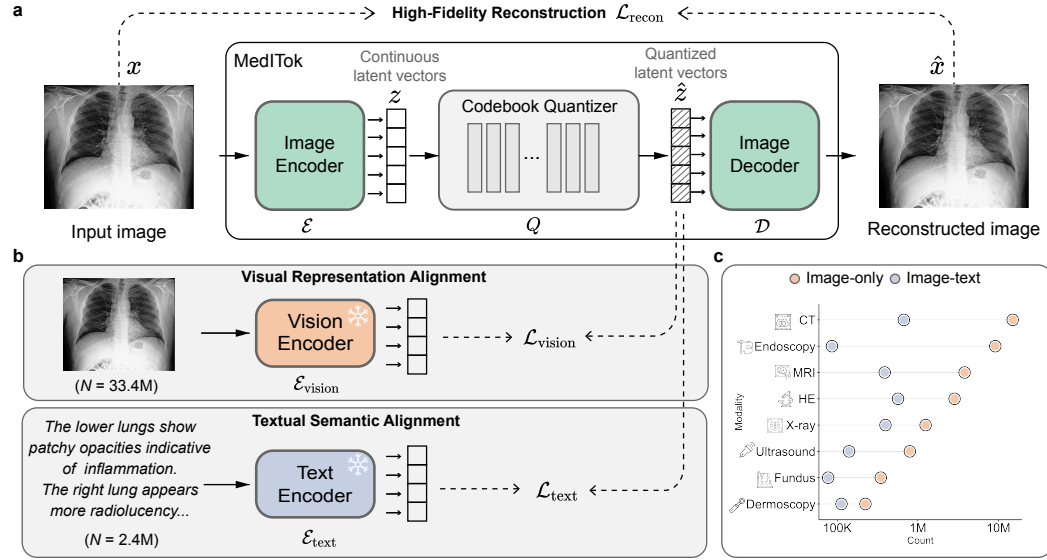


Figure 2: Overview of the proposed training framework. (a) Architecture of MedITok. (b) Two-stage training: visual representation alignment with pretrained visual semantics, followed by textual semantic alignment using clinical captions. (c) Statistics across modalities for our training data.

natural images, which encode insufficient clinical knowledge and capture either low-level structural detail or high-level clinical concepts, rarely both, whereas clinical tasks usually demand joint representation. To this end, we introduce MedITok, the first unified tokenizer tailored for medical images to support a wide range of tasks and empower advanced AR models in the medical field.

**Unified Visual Tokenizers.** Visual tokenizers encode images into token sequences suitable for AR modeling. Recent works (Wu et al., 2025; Ma et al., 2025b; Qu et al., 2024) seek to unify the encoding of both low-level details and high-level semantics into one single visual tokenizer, enhancing the multimodal generation and comprehension capabilities of downstream AR models. TokenFlow (Qu et al., 2024) proposes an intuitive dual-codebook design that explicitly decouples semantic and pixel-level cues. UniTok (Ma et al., 2025b) instead shows that simply scaling codebook capacity lets a single token space capture both, offering a more scalable solution. In medical imaging, recent research such as MedVAE (Varma et al., 2025) builds high-fidelity continuous latent compressors but stops short of providing unified tokens for downstream AR modeling. Our approach is the first medical-domain tokenizer to fill this gap, supplying unified token space to power the next generation of medical multimodal models.

### 3 METHODOLOGY

By encoding both low-level structural details and high-level clinical semantics, MedITok directly supports medical image reconstruction and classification tasks, and can be further integrated into AR models for more complex tasks, *e.g.*, medical image synthesis and interpretation, *etc.* Below, we start with a preliminary on the image tokenization (Sec. 3.1) and provide detailed description of our model and training framework (Sec. 3.2) and dataset curation process (Sec. 3.3).

#### 3.1 PRELIMINARY

The drive to apply powerful autoregressive models from natural language processing to visual data has spurred the development of image tokenization techniques, converting images into sequences of visual tokens. Among these, Vector Quantization (VQ)-based approaches (Van Den Oord et al., 2017; Esser et al., 2021) are foundational.

In a typical VQ-based image tokenizer, an image  $x$  is first mapped by an encoder  $\mathcal{E}$  to a spatial grid of latent vectors  $z \in \mathbb{R}^{h \times w \times d}$ . Each vector in this grid is then quantized by assigning it to the closest entry in a learned, finite codebook  $\mathcal{C} = \{c_k\}_{k=1}^K$ , where  $c_k \in \mathbb{R}^d$  represents a visual token and  $K$

is the codebook size. The quantized grid of latent vectors,  $z_q \in \mathbb{R}^{h \times w \times d}$ , effectively represent the image as a compressed sequence of visual tokens. A decoder  $\mathcal{D}$  is then trained to reconstruct the image from these representations, producing  $\hat{x} = \mathcal{D}(z_q)$ . During training, the encoder  $\mathcal{E}$ , decoder  $\mathcal{D}$ , and the codebook  $\mathcal{C}$  are jointly optimized. It typically involves a composite loss function designed to ensure both accurate reconstruction and effective codebook learning (Esser et al., 2021), defined as:

$$\mathcal{L}_{\text{recon}}(\hat{x}, x, z_q, z) = \mathcal{L}_{\text{image}}(\hat{x}, x) + \lambda_{\text{comm}} \mathcal{L}_{\text{comm}}(z_q, z), \quad (1)$$

where  $\mathcal{L}_{\text{image}}$  is the image fidelity loss consisting of a mean square error loss, a perceptual loss (Johnson et al., 2016), and an adversarial loss, encouraging high-fidelity reconstructions. The commitment loss (Van Den Oord et al., 2017)  $\mathcal{L}_{\text{comm}}$  ensures the encoder outputs  $z$  to commit to their nearest codebook vectors. Our work builds upon these foundational principles of VQ-based tokenization but introduces a novel training framework tailored to unified medical image tokenization.

### 3.2 MEDITOK TRAINING FRAMEWORK

A unified visual tokenizer must reconcile two objectives that naturally compete: preserving low-level spatial detail for image reconstruction and synthesis, and learning a high-level semantic token space for image interpretation. Previous works (Wu et al., 2025; Ma et al., 2025b) combine visual reconstruction and textual representation learning objectives in one go. Such training can lead to representation collapse or suboptimal trade-offs (Qu et al., 2024). Moreover, they typically rely on large-scale image-caption pairs while overlooking the abundance of unpaired images. We propose a novel two-stage training framework to train our unified visual tokenizer MedITok, unlocking the potential of unlabeled images in the medical domain and progressively transitioning from reconstruction-focused learning to unified token learning.

As depicted in Fig. 2, MedITok is comprised of an image encoder  $\mathcal{E}$ , a quantizer  $Q$ , and a decoder  $\mathcal{D}$ . Our framework begins with a *visual representation alignment* stage, which cold-starts the training of the image encoder  $\mathcal{E}$  and a decoder  $\mathcal{D}$  using a vast corpus of unpaired medical images. The primary focus is on capturing low-level structural information, guided by only a light semantic constraint from a pretrained vision encoder  $\mathcal{E}_{\text{vision}}$ . Subsequently, in the second stage termed *textual semantic alignment*,  $\mathcal{E}$  is refined using high-quality image-caption pairs. Here, the emphasis shifts towards enhancing the semantic richness of the learned tokens by aligning them with clinical captions processed by a pretrained text encoder  $\mathcal{E}_{\text{text}}$ . This approach not only alleviates the conflicts between reconstruction and semantic learning objectives but also allows us to effectively leverage large-scale unpaired images alongside paired image-text data for unified tokenizer training.

**Visual Representation Alignment.** Given an input image  $x$ , the encoder  $\mathcal{E}$  produces continuous latent vectors  $z$ , which are then quantized by the quantizer  $Q$  to yield discrete latent vectors  $z_q = Q(z)$ . The decoder  $\mathcal{D}$  then learns to reconstruct the image  $\hat{x} = \mathcal{D}(z_q)$ . The pretrained vision encoder  $\mathcal{E}_{\text{vision}}$  encodes the image  $x$  into a semantic representation, which is then projected into the space of  $z_q$  via a linear layer  $f_{\text{vision}}$  to provide semantic supervision for learning  $z_q$ . We use a composite loss function for training, defined as:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{recon}}(\hat{x}, x, z_q, z) + \lambda_{\text{vision}} \mathcal{L}_{\text{vision}}(z_q, f_{\text{vision}}(\mathcal{E}_{\text{vision}}(x))), \quad (2)$$

where  $\mathcal{L}_{\text{vision}}$  is a contrastive loss that imposes light semantic constraint on the latent space, with the factor  $\lambda_{\text{vision}}$  set to 0.1. By prioritizing reconstruction while gently guiding the latent space with pre-trained visual semantics, this stage ensures MedITok develops a robust understanding of visual structure, preparing it for fine-grained semantic alignment in the subsequent stage.

**Textual Semantic Alignment.** This stage focuses on enhancing the semantic richness of the learned image tokens and aligning them with fine-grained textual representations extracted from detailed medical captions. The training in this stage is driven by the following loss function:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{recon}}(\hat{x}, x, z_q, z) + \lambda_{\text{text}} \mathcal{L}_{\text{text}}(z_q, f_{\text{text}}(\mathcal{E}_{\text{text}}(t))), \quad (3)$$

where  $t$  denotes the detailed caption of the image  $x$ , and  $f_{\text{text}}$  is another linear layer.  $\mathcal{L}_{\text{text}}$  is the contrastive loss, with a balancing factor  $\lambda_{\text{text}}$  set to 1. This stage further integrates the structural and semantic representation learning, empowering MedITok for a wide range of downstream medical applications requiring nuanced understanding.



### 3.3 DATASET CURATION

The development of MedITok necessitates a comprehensive and diverse dataset. To this end, we undertake an extensive data collection effort, aggregating medical images and image-text pairs from over 300 publicly available sources. For example, image-text pairs are collected from BIOMED-ICA (Lozano et al., 2025), MedICaT (Subramanian et al., 2020), MIMIC-CXR (Johnson et al., 2019), ROCov2 (Rückert et al., 2024), PMC-OA (Lin et al., 2023), MM-Retinal (Wu et al., 2024), and GMAI-MM-Caption-1.7M (Li et al., 2024) datasets.

Quality control is a critical step in our data collection pipeline to ensure that the training data are of sufficient quality for learning meaningful representations. We employ a combination of automated and manual filtering to exclude images of low quality or limited clinical relevance. Specifically, an image is excluded if, after proxy RGB conversion, it meets any of the following criteria: (1) low pixel intensity range below 50; (2) insufficient resolution, where the smallest dimension is under 128 pixels; (3) low information content, indicated by a standard deviation of pixel values below 10; (4) limited palette, with three or fewer unique pixel values; (5) unrelated content, such as tables, plots, or non-clinical images extracted from publications. For text data, we only retain captions pertinent to clinical imaging, determined by the tags within each dataset or clinical keyword matching.

These checks efficiently remove noisy and uninformative samples and ensures higher quality input for our training framework, resulting in a massive corpus of 33,428,922 medical images for the visual representation alignment stage, and 2,422,827 high-quality medical image-caption pairs for the textual semantic alignment stage. This collection encompasses over eight imaging modalities, including computed tomography (CT), dermoscopy, endoscopy, fundus photography, magnetic resonance imaging (MRI), pathology, ultrasound, and X-ray, spanning a wide spectrum of anatomical regions and pathological findings. We leave detailed sources and statistics in our Appendix A.

## 4 EXPERIMENTS

In this section, we present comprehensive experiments to evaluate the proposed MedITok across four different task families, including medical image reconstruction, medical image classification, modality-conditioned medical image synthesis, and medical visual question answering. Since each task is evaluated using specialized metrics appropriate to its goals, we introduce them within each corresponding subsection.

### 4.1 EXPERIMENTAL SETUP

**Datasets.** (1) For medical image reconstruction, we collect images from 23 publicly available datasets (McCollough et al., 2017; Landman et al., 2015; Heimann et al., 2009; Kawahara et al., 2018; Giotis et al., 2015; Ali et al., 2022; Kiranyaz et al., 2020; Cartucho et al., 2024; Ali et al., 2020; Decenci re et al., 2014; Ovreiu et al., 2021; Fraz et al., 2012; Hoover et al., 2000; Graham et al., 2019b; Da et al., 2022; Nir et al., 2018b; Bao et al., 2025; Pati et al., 2020; Pedraza et al., 2015; Al-Dhabyani et al., 2020; Lian et al., 2021; Halabi et al., 2019; Tabik et al., 2020), totaling 35,736 images covering 8 imaging modalities. (2) For medical image classification, we benchmark on five subsets of the latest MedMNIST collection (Yang et al., 2023) in different imaging modalities, including BreastMNIST (Al-Dhabyani et al., 2020) for ultrasound, DermaMNIST (Tschandl et al., 2018; Codella et al., 2019) for dermoscopy, PathMNIST (Kather et al., 2019) for pathology images, PneumoniaMNIST (Kermay et al., 2018) for chest X-ray, and RetinaMNIST (Liu et al., 2022) for fundus photography, where all images are resized to  $256 \times 256$ . (3) For modality-conditioned medical image synthesis, we employ data from BloodMNIST (Acevedo et al., 2020), BreastMNIST, ChestMNIST (Wang et al., 2017b), DermaMNIST, PathMNIST, and RetinaMNIST to train and test the downstream AR image synthesis models. (4) Finally, for medical visual question answering, we use PubMedVision (Chen et al., 2024c) dataset to train the downstream multimodal language models, and evaluate them on two widely used medical visual question answering (VQA) benchmarks: VQA-RAD (Lau et al., 2018) and SLAKE (Liu et al., 2021a). We carefully conduct manual cross-checking on the data used for evaluating and training MedITok, minimizing the risk of data leakage. Please see Appendix B for more details on statistics and tasks.

**Implementation Detail.** We build MedITok with a hybrid ViT architecture (Chen et al., 2024b) using PyTorch (Paszke et al., 2019), and implement the quantizer with 8 codebooks, each containing

Table 1: Medical image reconstruction across different imaging modalities using different models. The best results are highlighted in **bold** and the second best results are underlined. SSIM values are presented as percentages.  $f_d$  denotes the downsampling factor. “↓”: The lower the better.

Metrics	Models	$f_d$	CT	Dermo.	Endo.	Fundus.	MRI	Patho.	US	X-ray	Avg.	Avg. rank
rFID (↓)	VQGAN	8	15.97	33.57	27.33	27.22	21.33	67.68	29.48	18.66	30.16	4.9
	Emu3-VQ	8	11.83	27.91	20.83	16.27	13.52	69.89	25.43	11.99	24.71	3.4
	VAR-VQ	16	14.69	30.27	19.74	21.69	13.99	70.06	26.09	12.18	26.09	4.1
	TokenFlow	16	24.78	44.28	47.42	34.93	26.81	98.22	51.77	24.51	44.09	7.0
	UniTok	16	<u>9.27</u>	<u>23.15</u>	<u>13.64</u>	<u>16.22</u>	<u>9.30</u>	<u>47.77</u>	<u>20.93</u>	<u>8.61</u>	<u>18.61</u>	<u>2.0</u>
	PUMIT	16	32.67	53.46	56.22	27.51	25.43	142.98	37.04	23.78	49.88	7.1
	MedVAE	8	20.17	140.39	114.00	117.39	23.34	123.20	30.60	11.54	73.64	6.5
	MedITok	16	<b>7.88</b>	<b>22.27</b>	<b>10.66</b>	<b>14.39</b>	<b>6.32</b>	<b>46.54</b>	<b>17.64</b>	<b>6.55</b>	<b>16.53</b>	<b>1.0</b>
	VQGAN	8	31.13	29.28	25.60	35.40	29.54	20.42	24.79	31.68	28.48	6.3
	Emu3-VQ	8	36.11	<u>31.68</u>	28.96	<b>39.64</b>	<u>34.32</u>	22.08	27.57	<u>35.81</u>	<b>32.02</b>	2.6
PSNR (↑)	VAR-VQ	16	31.32	29.26	25.75	35.73	29.83	20.86	25.22	31.10	28.63	5.8
	TokenFlow	16	28.64	27.23	23.72	33.45	27.68	19.33	23.26	28.71	26.50	7.8
	UniTok	16	33.60	30.97	27.55	37.21	31.50	22.18	26.97	32.97	30.34	4.3
	PUMIT	16	33.64	30.23	29.08	37.33	33.13	<u>23.09</u>	28.31	33.89	31.09	3.1
	MedVAE	8	<b>36.46</b>	20.67	25.04	15.31	<b>34.42</b>	19.58	<u>28.29</u>	<b>36.23</b>	27.01	4.5
	MedITok	16	<u>36.32</u>	<b>31.69</b>	<b>29.19</b>	<u>37.72</u>	33.55	<b>23.54</b>	<b>28.49</b>	34.42	<u>31.74</u>	<b>1.8</b>
	VQGAN	8	88.51	75.28	76.84	92.32	84.39	48.42	68.18	91.14	78.14	6.8
	Emu3-VQ	8	92.79	79.34	<u>84.71</u>	94.33	95.72	54.70	75.14	<b>95.54</b>	83.78	3.5
	VAR-VQ	16	89.51	76.69	79.21	93.08	93.68	47.40	69.99	90.79	80.04	6.0
	TokenFlow	16	82.43	67.19	69.47	89.60	90.22	33.09	56.56	84.50	71.63	7.8
SSIM (↑)	UniTok	16	92.42	<u>81.00</u>	84.47	<u>94.45</u>	95.47	56.42	76.40	92.74	84.17	3.9
	PUMIT	16	92.10	85.41	87.81	94.60	96.59	<u>63.81</u>	<u>81.46</u>	94.52	<u>87.04</u>	2.6
	MedVAE	8	<u>92.86</u>	75.32	81.52	69.46	<u>95.92</u>	53.10	77.45	94.77	80.10	4.4
	MedITok	16	<b>93.73</b>	<b>85.47</b>	<b>88.99</b>	<b>95.27</b>	<b>97.22</b>	<b>65.99</b>	<b>83.93</b>	<u>95.39</u>	<b>88.25</b>	<b>1.1</b>

4,096 eight-dimensional latent vectors. We train MedITok using AdamW (Loshchilov & Hutter, 2019) optimizer for 3 epochs in the first stage and 2 epochs in the second stage, with a global batch size of 512. Image resolution is of  $256 \times 256$ . The encoder of MedITok is initialize with weights from UniTok for efficient training. We choose BiomedCLIP (Zhang et al., 2023b) as the pretrained semantic vision and text encoders for alignment in our training framework, which is frozen throughout the training. Detailed setup can be found in our Appendix C.

**Competing Tokenizers.** We compare MedITok with powerful visual tokenizers from both general and medical domains, including VQGAN (Esser et al., 2021), Emu3-VQ (Wang et al., 2024b), VAR-VQ (Tian et al., 2024), TokenFlow (Qu et al., 2024), UniTok (Ma et al., 2025b), PUMIT (Luo et al., 2023b), and MedVAE (Varma et al., 2025). VQGAN, Emu3-VQ, and VAR-VQ are pure VQ-based tokenizers trained on natural images without semantic alignments, yet showing great promise in building medical multimodal models (Lin et al., 2025; Ma et al., 2025a). TokenFlow and UniTok are two state-of-the-art unified image tokenizers in the general domains. PUMIT and MedVAE are two medical visual tokenizers that mainly focus on fine-grained detail preservation.

## 4.2 MEDICAL IMAGE RECONSTRUCTION

We employ reconstruction Fréchet inception distance (rFID) (Heusel et al., 2017), peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM) (Wang et al., 2004) to evaluate the image reconstruction performance. Notably, Woodland et al. (2024) have shown that ImageNet-pretrained feature extractors are more consistent and aligned with human medical expert judgment than their counterparts pretrained on medical images, and we follow their work to implement rFID.

Quantitative results are shown in Table 1. MedVAE struggles on the modalities with colored imaging (e.g., fundus photography) as it is trained only on grayscale images (Varma et al., 2025). Notably, despite with a large downsampling factor of  $16\times$ , MedITok delivers the best overall reconstruction quality across 8 modalities, outperforming other tokenizers including those with only  $8\times$  downsampling. This highlights the efficiency of MedITok in balancing compression with reconstruction fidelity. Fig. 3 visualizes images reconstructed by different tokenizers and corresponding error maps. MedVAE fails to preserve colors due to limited generalizability, while UniTok discards nuanced details. By contrast, our MedITok consistently preserves fine-grained structures and color fidelity. Please refer to Appendix D for more results.

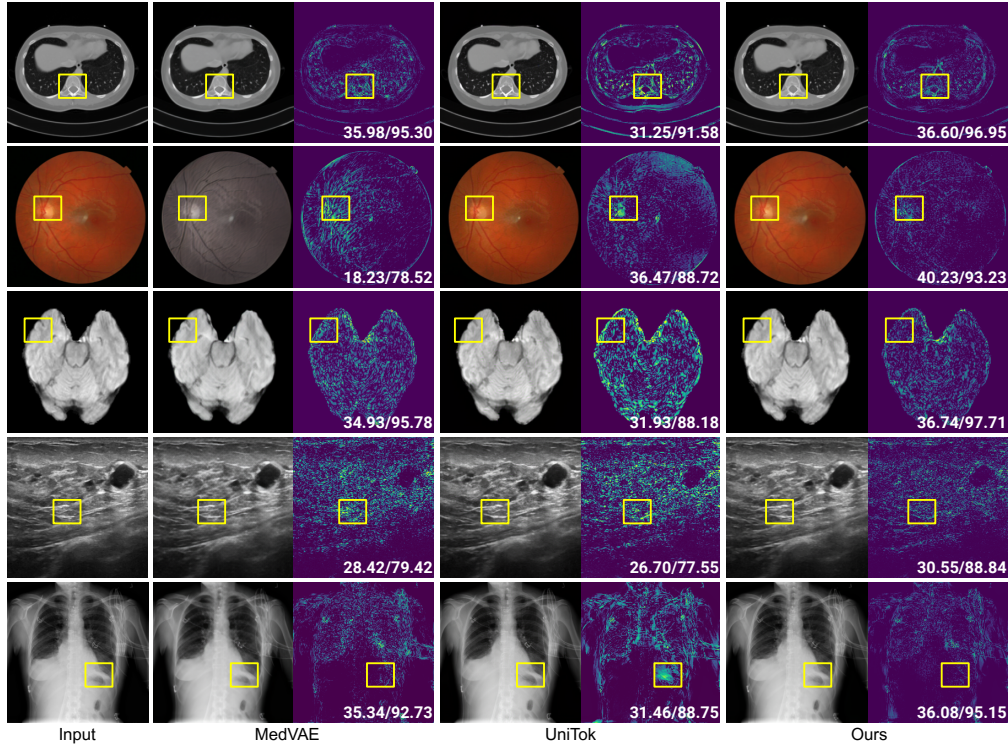


Figure 3: Reconstruction results across multiple imaging modalities. Each reconstructed image is paired with an absolute error map against the input image with PSNR/SSIM values.

Table 2: Downstream image classification performance (mAP / AUC) with linear probing setup. The best results are highlighted in **bold** and the second best results are underlined. Values are presented as percentages.

Models	Dermoscopy	Fundus	Pathology	Ultrasound	X-ray	AVG
VQGAN	35.71/85.97	41.59/77.33	72.69/94.57	73.29/76.35	91.34/93.32	62.92/85.51
Emu3-VQ	30.79/82.88	38.90/71.71	42.57/82.75	82.65/85.30	92.75/93.29	57.53/83.19
VAR-VQ	58.76/94.02	51.71/85.53	90.80/98.31	87.31/89.06	97.56/97.79	77.23/92.94
TokenFlow	61.78/93.50	52.07/83.77	95.21/99.23	<b>88.19/88.12</b>	<u>97.69/98.03</u>	78.99/92.53
UniTok	<u>66.16/94.60</u>	<u>55.94/85.05</u>	<u>96.63/99.49</u>	87.34/88.60	95.98/96.84	<u>80.41/92.92</u>
PUMIT	23.64/71.92	36.60/72.87	81.52/96.50	68.81/73.67	88.80/91.64	59.87/81.31
MedVAE	37.66/85.26	39.31/75.29	48.02/84.85	77.74/82.36	95.41/95.47	59.54/84.64
MedITok (ours)	<b>71.52/95.60</b>	<b>56.41/86.88</b>	<b>96.88/99.60</b>	<u>87.45/89.07</u>	<b>99.08/99.19</b>	<b>82.27/94.07</b>

### 4.3 MEDICAL IMAGE CLASSIFICATION

To assess whether a visual tokenizer encodes clinical semantics in the latent space, we adopt a linear-probing (Alain & Bengio, 2016) protocol on a suite of medical image classification tasks (Yang et al., 2023). Specifically, we freeze each tokenizer and append a linear layer to its encoder, training the linear layer to convergence on the image classification task and report the performance in terms of mean average precision (mAP) and area under the ROC curve (AUC) on the corresponding test sets. Results are presented in Table 2. Models optimized purely for image reconstruction (e.g., Emu3-VQ, PUMIT) degrade on tasks where fine-grained clinical semantics are required for nuanced classification, e.g., retinal disease grading in fundus photographs. General-domain unified tokenizers like TokenFlow and UniTok show improved but limited performance. By contrast, our MedITok encodes rich clinical-relevant semantics and delivers the best overall classification performance, showing that rich semantic information is embedded in MedITok’s unified token space.



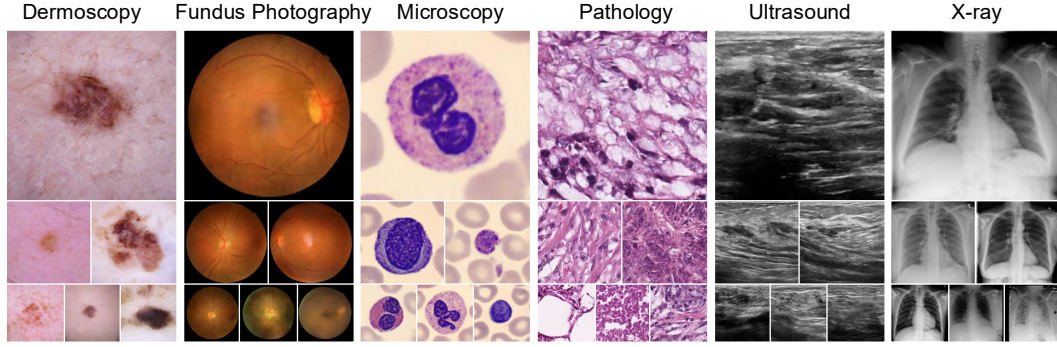
Figure 4: Modality-conditioned synthesized image examples produced by our LlamaGen<sub>MedITok</sub>.

Table 3: Modality-conditioned medical image generation performance.

Models	gFID ( $\downarrow$ )	Diversity ( $\uparrow$ )
LlamaGen <sub>VQGAN</sub>	130.93 $\pm$ 3.58	0.6503 $\pm$ 0.03
LlamaGen <sub>UniTok</sub>	80.71 $\pm$ 3.18	0.6584 $\pm$ 0.02
LlamaGen <sub>MedITok-S1</sub>	94.98 $\pm$ 1.89	0.6479 $\pm$ 0.02
<b>LlamaGen<sub>MedITok</sub></b>	<b>76.78<math>\pm</math>1.91</b>	<b>0.6883<math>\pm</math>0.01</b>

Table 4: Visual question answering accuracy on two medical benchmarks.

Models	VQARAD	SLAKE-val	SLAKE-test
LLaVA-Med	43.90 $\pm$ 2.88	40.30 $\pm$ 2.28	38.73 $\pm$ 3.53
LLaVA-Med <sub>UniTok</sub>	49.66 $\pm$ 1.11	44.44 $\pm$ 2.04	43.84 $\pm$ 1.28
LLaVA-Med <sub>MedITok-S1</sub>	46.56 $\pm$ 1.67	40.73 $\pm$ 1.52	41.02 $\pm$ 0.83
<b>LLaVA-Med<sub>MedITok</sub></b>	<b>52.99<math>\pm</math>2.14</b>	<b>49.02<math>\pm</math>3.45</b>	<b>48.09<math>\pm</math>1.42</b>

#### 4.4 MEDICAL IMAGE SYNTHESIS

We explore applying unified visual tokenizers to image synthesis task by incorporating each tokenizer in the LlamaGen (Sun et al., 2024) framework for modality-conditioned medical image synthesis, including six imaging modalities: dermoscopy, fundus photography, microscopy, pathology images, ultrasound, and X-ray. Specifically, we build two LlamaGen models using MedITok-S1, a variant of MedITok that only goes through the first training stage, and MedITok. These two models, denoted by “LlamaGen<sub>MedITok-S1</sub>” and “LlamaGen<sub>MedITok</sub>”, respectively, are compared with other LlamaGen variants with different visual tokenizers, *i.e.*, “LlamaGen<sub>VQGAN</sub>” and “LlamaGen<sub>UniTok</sub>”. We follow previous work (Bluthgen et al., 2024) to report generation Fréchet inception distance (gFID) and diversity score for evaluating the fidelity and the diversity of the synthesized images. For visual diversity metric, we first sample  $N$  images from the modality-to-image model for each modality, and calculate  $\sum_{i \neq j} (1 - \text{MS-SSIM}(x_i, x_j)) / \binom{N}{2}$  for all distinct pairs  $(x_i, x_j)$  among  $N$  synthesized images, where MS-SSIM denotes the multi-scale structural similarity index (Wang et al., 2003). The overall diversity score is defined as the mean diversity score over all imaging modalities.

Quantitative results in Table 3 show that LlamaGen using general-domain tokenizer like VQ-GAN or UniTok struggles with high-quality medical image generation. Notably, LlamaGen<sub>MedITok</sub> achieves the best visual fidelity and diversity. We also note that LlamaGen<sub>MedITok</sub> surpasses LlamaGen<sub>MedITok-S1</sub> by a non-trivial margin, indicating that textual semantic alignment may regularize the token space and boost the image synthesis task. Fig. 4 illustrates images synthesized by LlamaGen<sub>MedITok</sub> across various modalities, presenting realistic structures and textures of biological tissues and organs. Note that, although MedITok is not trained on microscopy modalities, it still supports realistic synthesis of microscopy images. Please refer to Appendix D for more examples.

#### 4.5 MEDICAL IMAGE INTERPRETATION

We further evaluate the effectiveness of different tokenizers in medical image interpretation by integrating each as the image encoder in the LLaVA-Med (Li et al., 2023) framework, yielding three models: LLaVA-Med<sub>UniTok</sub>, LLaVA-Med<sub>MedITok-S1</sub>, and LLaVA-Med<sub>MedITok</sub>. We initialize the language backbone using the released weights of LLaVA-Med, train these models on the PubMed-Vision (Chen et al., 2024c) dataset, and evaluate their accuracy on two widely used medical VQA benchmarks: VQA-RAD (Lau et al., 2018) and SLAKE (Liu et al., 2021a).

Table 5: Ablation studies of MedITok. “#Img”: number of images used in the first training stage, “#Img-txt”: number of image-text pairs used in the second training stage.

Idx.	Vision Target Repr.	Text Target Repr.	Objective	#Img	#Img-txt	PSNR	SSIM	mAP	AUC
(i)	CLIP-V	–	Contrast	800k	–	30.99	86.67	70.80	89.01
(ii)	BiomedCLIP-V	–	Contrast	800k	–	30.00	83.85	78.35	92.23
(iii)	BiomedCLIP-V	BiomedCLIP-T	Contrast	800k	1M	30.03	84.32	80.09	92.64
(iv)	BiomedCLIP-V	–	Contrast	1.8M	–	31.38	84.36	78.49	92.25
(v)	BiomedCLIP-V	BiomedCLIP-T	Contrast	800k	2.4M	29.74	84.14	80.28	92.72
(vi)	BiomedCLIP-V	BiomedCLIP-T	Contrast	2M	2.4M	30.20	85.50	82.23	93.61
(vii)	BiomedCLIP-V	BiomedCLIP-T	Contrast	33.4M	2.4M	31.74	88.25	82.27	94.07

As shown in Table 4, LLaVA-Med equipped with our final MedITok consistently outperforms other variants across all benchmarks. We observe a similar improvement from MedITok-S1 to MedITok as in Table 2, indicating the necessity of the textual semantic alignment stage. The underperformance of general-domain tokenizer, UniTok, highlights the importance of domain-specific semantic encoding. These results confirm that MedITok provides effective representations to develop powerful AR models for downstream medical image interpretation tasks.

#### 4.6 ABLATION STUDIES

**Choice of Pre-trained Encoder.** Ideally, the pretrained encoders in the proposed training framework are designed to be flexible, provided they offer rich semantic representations, *e.g.*, CLIP-family (Radford et al., 2021; Zhang et al., 2023b). We experiment with two options: the general-domain CLIP (Radford et al., 2021) and the biomedical-domain BiomedCLIP (Zhang et al., 2023b). Results in Rows (i) and (ii) of Table 5 show that, by aligning to the representations produced from BiomedCLIP, MedITok achieves significant improvement in the medical image classification tasks while maintaining competitive image reconstruction performance, indicating that domain-specific pre-trained encoders can provide clinically-relevant semantics that benefit downstream medical tasks.

**Two-Stage Training.** We further validate the importance of the textual semantic alignment stage by comparing our full two-stage framework, shown in Row (iii) of Table 5, against a single-stage variant with the same number of training images, shown in Row (iv). Our two-stage approach boosts image classification without degrading reconstruction quality, highlighting that the textual representation alignment stage helps the model capture richer cross-modal semantics.

**Image Data Scaling.** One notable benefit of our proposed training framework is that it allows effective use of unpaired medical image datasets, which are typically more accessible than image-text data. Rows (v), (vi), and (vii) of Table 5 illustrate the impact of scaling up the number of unpaired image corpus in the first training stage. Notably, expanding the image data from 800k to 33.4M yields consistent improvements across all metrics, demonstrating the scalability of our proposed approach, allowing it to fully exploit medical image data to enhance both structural fidelity and downstream diagnostic performance. More experimental results can be found in Appendix D.

## 5 CONCLUSION

In this paper, we propose MedITok, a unified medical image tokenizer that encodes both low-level structural details and high-level clinical semantics. Leveraging a novel two-stage training framework which involves visual representation alignment on large-scale unpaired images and textual semantic alignment on high-quality image-caption pairs, MedITok learns a unified token space that facilitates medical image reconstruction, classification, synthesis, and VQA, outperforming existing general-domain and medical-domain models across multiple imaging modalities. By providing a unified token space, we believe MedITok will serve as a foundational building block for next-generation multimodal models in medical applications. Please refer to Appendix E for more discussion.



**Ethics Statement.** We affirm adherence to the ICLR Code of Ethics. This work uses only publicly available datasets with clear licensing; no new human or animal subjects were recruited and no protected health information beyond what is already de-identified in the source data was used. We discuss potential societal risks in Appendix E.4, including bias, misuse of generative models, and the need for oversight. Large language models were used only to aid and polish wording, improving the flow and clarity of the presentation; they did NOT generate analyses, experiments, figures, or results, and all technical content was authored by the authors.

**Reproducibility Statement.** We provide anonymized code and access to model weights in our supplementary material. Training data sources, preprocessing, and statistics are detailed in Appendix A with dataset lists (Tables S5–S8). Evaluation datasets, task definitions, and metrics appear in Sec. 4 and Appendix B (including Tables S10 and S11). Experimental setups are detailed in Sec. 4 and Appendix C.

## REFERENCES

- Aptos 2019 blindness detection. <https://www.kaggle.com/competitions/aptos2019-blindness-detection>. Accessed: 2025-05-22.
- Machine learning for pseudopapilledema, 2020. URL <https://osf.io/2w5ce/>. Accessed: 2025-05-22.
- Adam Abeshouse, Jaeil Ahn, Rehan Akbani, Adrian Ally, Samirkumar Amin, Christopher D Andry, Matti Annala, Armen Aprikian, Joshua Armenia, Arshi Arora, et al. The molecular taxonomy of primary prostate cancer. *Cell*, 163(4):1011–1025, 2015.
- Nurmukhammed Baktybekuly Abuev. Augmented ocular diseases [data set], 2021. URL <https://www.kaggle.com/datasets/nurmukhammed7/augmented-ocular-diseases>. Accessed: 2025-05-22.
- Andrea Acevedo, Anna Merino, Santiago Alf  rez,   ngel Molina, Laura Bold  , and Jos   Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30:105474, 2020.
- Maruf Adewole, Jeffrey D. Rudie, Anu Gbadamosi, Oluyemisi Toyobo, Confidence Raymond, Dong Zhang, Olubukola Omidiji, Rachel Akinola, Mohammad Abba Suwaid, Adaobi Emegoakor, Nancy Ojo, Kenneth Aguh, Chinasa Kalaiwo, Gabriel Babatunde, Afolabi Ogunleye, Yewande Gbadamosi, Kator Iorpagher, Evan Calabrese, Mariam Aboian, Marius Linguraru, Jake Albrecht, Benedikt Wiestler, Florian Kofler, Anastasia Janas, Dominic LaBella, Anahita Fathi Kzerooni, Hongwei Bran Li, Juan Eugenio Iglesias, Keyvan Farahani, James Eddy, Timothy Bergquist, Verena Chung, Russell Takeshi Shinohara, Walter Wiggins, Zachary Reitman, Chunhao Wang, Xinyang Liu, Zhifan Jiang, Ariana Familiar, Koen Van Leemput, Christina Bukas, Maire Piraud, Gian-Marco Conte, Elaine Johansson, Zeke Meier, Bjoern H Menze, Ujjwal Baid, Spyridon Bakas, Farouk Dako, Abiodun Fatade, and Uduinna C Anazodo. The brain tumor segmentation (brats) challenge 2023: Glioma segmentation in sub-saharan africa patient population (brats-africa). *arXiv preprint arXiv:2305.19369*, 2023.
- Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, Ren   Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1):4006, 2014.
- Oguz Akin, Pierre Elnajjar, Matthew Heller, Rose Jarosz, Bradley J. Erickson, Shanah Kirk, Yueh Lee, Marston W. Linehan, Rabindra Gautam, Raghu Vikram, Kimberly M. Garcia, Charles Roche, Ermelinda Bonaccio, and Joe Filippini. The cancer genome atlas kidney renal clear cell carcinoma collection (tcga-kirc), 2016. URL <https://www.cancerimagingarchive.net/collection/tcga-kirc/>.
- Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020.

- Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.
- Syed Samiul Alam, Samiul Based Shuvo, Shams Nafisa Ali, Fardeen Ahmed, Arbil Chakma, and Yeong Min Jang. Benchmarking deep learning frameworks for automated diagnosis of ocular toxoplasmosis: A comprehensive approach to classification and segmentation. *arXiv preprint arXiv:2305.10975*, 2023.
- 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>.
- Shams Nafisa Ali, Md Tazuddin Ahmed, Joydip Paul, Tasnim Jahan, SM Sani, Nawsabah Noor, and Taufiq Hasan. Monkeypox skin lesion detection using deep learning models: A feasibility study. *arXiv preprint arXiv:2207.03342*, 2022.
- Sharib Ali, Noha Ghatwary, Barbara Braden, Dominique Lamarque, Adam Bailey, Stefano Realdon, Renato Cannizzaro, Jens Rittscher, Christian Daul, and James East. Endoscopy disease detection challenge 2020. *arXiv preprint arXiv:2003.03376*, 2020.
- Ahmed Almazroa, Sami Alodhayb, Essameldin Osman, Eslam Ramadan, Mohammed Hummadi, Mohammed Dlaim, Muhannad Alkatee, Kaamran Raahemifar, and Vasudevan Lakshminarayanan. Retinal fundus images for glaucoma analysis: the RIGA dataset. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, volume 10579, pp. 55–62. SPIE, 2018.
- Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai AT Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem SE Salem, Ahmed F Ismail, Anas M Saad, et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18): 3461–3467, 2019.
- MD Anouk Stein, Carol Wu, Chris Carr, George Shih, Jamie Dulkowski, kalpathy, Leon Chen, Luciano Prevedello, MD Marc Kohli, Mark McDonald, Peter, Phil Culliton, Safwan Halabi MD, and Tian Xia. Rsn pneumonia detection challenge. <https://kaggle.com/competitions/rsna-pneumonia-detection-challenge>, 2018. Kaggle.
- 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.
- 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.
- 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.
- Samuel G. Armato III, 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://www.cancerimagingarchive.net/collection/spie-aapm-lung-ct-challenge/>.
- ZI Amanullah Asraf and Zahirul Islam. Covid19, pneumonia and normal chest x-ray pa dataset, 2021.
- Itzik Avital, Ilya Nelkenbaum, Galia Tsarfaty, Eli Konen, Nahum Kiryati, and Arnaldo Mayer. Neural segmentation of seeding rois (srois) for pre-surgical brain tractography. *IEEE Transactions on Medical Imaging*, 39(5):1655–1667, 2019.

- Bhakti Baheti, Satrajit Chakrabarty, Hamed Akbari, Michel Bilello, Benedikt Wiestler, Julian Schwarting, Evan Calabrese, Jeffrey Rudie, Syed Abidi, Mina Mousa, et al. The brain tumor sequence registration (brats-reg) challenge: Establishing correspondence between pre-operative and follow-up mri scans of diffuse glioma patients. *arXiv preprint arXiv:2112.06979*, 2021.
- Baidu AI Studio. X-ray hand joint classification dataset [data set], 2021. URL <https://aistudio.baidu.com/datasetdetail/69582/0>. Accessed: 2025-05-22.
- Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels for the pre-operative scans of the tcga-lgg collection, 2017a. URL <https://www.cancerimagingarchive.net/analysis-result/brats-tcga-lgg/>.
- Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels for the pre-operative scans of the tcga-gbm collection, 2017b. URL <https://www.cancerimagingarchive.net/analysis-result/brats-tcga-gbm/>.
- Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4(1):1–13, 2017c.
- 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.
- Sophia Bano, Francisco Vasconcelos, Luke M Shepherd, Emmanuel Vander Poorten, Tom Vercauteren, Sebastien Ourselin, Anna L David, Jan Deprest, and Danail Stoyanov. Deep placental vessel segmentation for fetoscopic mosaicking. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 763–773, 2020.
- Rina Bao, Ya’nan Song, Sara V Bates, Rebecca J Weiss, Anna N Foster, Camilo Jaimes, Susan Sotardi, Yue Zhang, Randy L Hirschtick, P Ellen Grant, et al. Boston neonatal brain injury data for hypoxic ischemic encephalopathy (bonbid-hie): I. mri and lesion labeling. *Scientific Data*, 12(1): 53, 2025.
- Francisco José Fumero Batista, Tinguaro Diaz-Aleman, Jose Sigut, Silvia Alayon, Rafael Arnay, and Denisse Angel-Pereira. Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning. *Image Analysis & Stereology*, 39(3):161–167, 2020.
- 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>.
- Veronica Elisa Castillo Benítez, Ingrid Castro Matto, Julio César Mello Román, José Luis Vázquez Noguera, Miguel García-Torres, Jordan Ayala, Diego P. Pinto-Roa, Pedro E. Gardel-Sotomayor, Jacques Facon, and Sebastian Alberto Grillo. Dataset from fundus images for the study of diabetic retinopathy, February 2021.
- 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.
- Cancer Moonshot Biobank. Cancer moonshot biobank – gastroesophageal cancer collection (cmb-gec), 2022a. URL <https://www.cancerimagingarchive.net/collection/cmb-gec/>. Accessed: 2025-05-22.
- Cancer Moonshot Biobank. Cancer moonshot biobank – melanoma collection (cmb-mel), 2022b. URL <https://www.cancerimagingarchive.net/collection/cmb-mel/>. Accessed: 2025-05-22.

- Christian Bluethgen, Pierre Chambon, Jean-Benoit Delbrouck, Rogier van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay S Chaudhari. A vision–language foundation model for the generation of realistic chest x-ray images. *Nature Biomedical Engineering*, pp. 1–13, 2024.
- Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, Dag Johansen, Carsten Griwodz, Håkon K Stensland, Enrique Garcia-Ceja, Peter T Schmidt, Hugo L Hammer, Michael A Riegler, Pål Halvorsen, and Thomas de Lange. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283, 2020.
- Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019.
- 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>.
- Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, Maria Gabrani, Florinda Feroce, and Maria Frucci. BRACS: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022:baac093, 10 2022. ISSN 1758-0463. doi: 10.1093/database/baac093.
- Attila Budai, Rüdiger Bock, Andreas Maier, Joachim Hornegger, and Georg Michelson. Robust vessel segmentation in fundus images. *International Journal of Biomedical Imaging*, 2013(1): 154860, 2013.
- Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester Van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, 28 (1):154–163, 2022.
- Myrthe AD Buser, Dominique C Simons, Matthijs Fitski, MHW A Wijnjen, Annemieke S Littooi, Annemiek H ter Brugge, Iris N Vos, Markus HA Janse, Mathijs de Boer, Rens ter Maat, et al. Automated segmentation of pediatric neuroblastoma on multimodal mri: Results of the sppin challenge at miccai 2023. *arXiv preprint arXiv:2505.00369*, 2025.
- Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.
- Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P. Musco, Rahul K. Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J. McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33(2):577–590, 2014a.
- Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P. Musco, Rahul K. Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J. McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33(2):577–590, 2014b.
- Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017.
- Carlos Cardenas, Abdallah Mohamed, Greg Sharp, Mark Gooding, Harini Veeraraghavan, and Yang Jinzhong. Data from aapm rt-mac grand challenge 2019, 2019. URL <https://www.cancerimagingarchive.net/collection/aapm-rt-mac/>.

- Olivia Cardozo, Verena Ojeda, Rodrigo Parra, Julio César Mello-Román, José Luis Vázquez Noguera, Miguel García-Torres, Federico Divina, Sebastian A Grillo, Cynthia Villalba, Jacques Facon, et al. Dataset of fundus images for the diagnosis of ocular toxoplasmosis. *Data in Brief*, 48:109056, 2023.
- João Cartucho, Alistair Weld, Samyakh Tukra, Haozheng Xu, Hiroki Matsuzaki, Taiyo Ishikawa, Minjun Kwon, Yong Eun Jang, Kwang-Ju Kim, Gwang Lee, et al. SurgT challenge: Benchmark of soft-tissue trackers for robotic surgery. *Medical Image Analysis*, 91:102985, 2024.
- 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.
- Pierre Chambon, Jean-Benoît 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.
- E Chen, Hua-Nong Ting, Joon-Huang Chuah, and Jun Zhao. Segmentation of overlapping cells in cervical cytology images: a survey. *IEEE Access*, 2024a.
- 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.
- Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Vitamin: Designing scalable vision models in the vision-language era. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12954–12966, 2024b.
- 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*, 2024c.
- Pingjun Chen. Knee osteoarthritis severity grading dataset, 2018. URL <https://doi.org/10.17632/56rmx5bjcr.1>.
- Tao Chen, Chenhui Wang, Zhihao Chen, and Hongming Shan. Autoregressive medical image segmentation via next-scale mask prediction. *arXiv preprint arXiv:2502.20784*, 2025a.
- Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.
- Stephanie J Chiu, Yuliya Lokhnygina, Adam M Dubis, Alfredo Dubra, Joseph Carroll, Joseph A Izatt, and Sina Farsiu. Automatic cone photoreceptor segmentation using graph theory and dynamic programming. *Biomedical optics express*, 4(6):924–937, 2013.
- Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8: 132665–132676, 2020.
- Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019.
- Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), 2018a. URL <https://arxiv.org/abs/1710.05006>.



- Noel C. F. Codella, David A. Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kallou, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172, 2018b.
- Olivier Commowick, Audrey Istace, Michael Kain, Baptiste Laurent, Florent Leray, Mathieu Simon, Sorina Camarasu Pop, Pascal Girard, Roxana Ameli, Jean-Christophe Ferré, et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific Reports*, 8(1):13650, 2018.
- National Cancer Institute Clinical Proteomic Tumor Analysis Consortium et al. The clinical proteomic tumor analysis consortium pancreatic ductal adenocarcinoma collection (cptac-pda). *The Cancer Imaging Archive* <https://doi.org/10.7937/K.9,2018>.
- Chunyan Cui, Li Li, Hongmin Cai, Zhihao Fan, Ling Zhang, Tingting Dan, Jiao Li, and Jinghua Wang. The chinese mammography database (cmmd): An online mammography database with biopsy confirmed types for machine diagnosis of breast, 2021. URL <https://doi.org/10.7937/tcia.eqde-4b16>. Accessed: 2025-05-22.
- Qian Da, Xiaodi Huang, Zhongyu Li, Yanfei Zuo, Chenbin Zhang, Jingxin Liu, Wen Chen, Jiahui Li, Dou Xu, Zhiqiang Hu, et al. DigestPath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Medical Image Analysis*, 80:102485, 2022.
- B Dashtbozorg, J Zhang, F Huang, and BM ter Haar Romenij. Retinacheck: an interactive platform for retinal image analysis. In *13th IEEE International Symposium on Biomedical Imaging (ISBI 2016)*, 2016.
- Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, Adrian Galdran, Miguel Ángel González Ballester, Gustavo Carneiro, R. G. Devika, Hrishikesh Panikkasseril Sethumadhavan, Densen Puthussery, Hong Liu, Zekang Yang, Satoshi Kondo, Satoshi Kasai, Edward Wang, Ashritha Durvasula, Jónathan Heras, Miguel Ángel Zapata, Teresa Araújo, Guilherme Aresta, Hrvoje Bogunović, Mustafa Arikani, Yeong Chan Lee, Hyun Bin Cho, Yoon Ho Choi, Abdul Qayyum, Imran Razzak, Bram van Ginneken, Hans G. Lemij, and Clara I. Sánchez. AIROGS: Artificial intelligence for robust glaucoma screening challenge. *IEEE Transactions on Medical Imaging*, 43(1):542–557, 2024.
- Etienne Decenciere, Guy Cazuguel, Xiwei Zhang, Guillaume Thibault, J-C Klein, Fernand Meyer, Beatriz Marcotegui, Gwénolé Quéllec, Mathieu Lamard, Ronan Danno, et al. Teleophtha: Machine learning and image processing methods for teleophthalmology. *Irbm*, 34(2):196–203, 2013.
- Etienne Decencièrè, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, John-Richard Ordóñez-Varela, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, pp. 231–234, 2014.
- 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.
- Derbi Hackathon Organizers. Derbi hackathon retinal fundus image dataset. <https://www.kaggle.com/datasets/nikkich9/derbi-hackathon-retinal-fundus-image-dataset>, 2022. Accessed: 2025-05-22.
- Shivang Desai, Ahmad Baghal, Thidathip Wongsurawat, Shaymaa Al-Shukri, Kim Gates, Phillip Farmer, Michael Rutherford, Geri D Blake, Tracy Nolan, Thomas Powell, et al. Chest imaging with clinical and genomic correlates representing a rural covid-19 positive population. (*No Title*), 2020.

- 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.
- Emma Dugas, Jared, Jorge, and Will Cukierski. Diabetic retinopathy detection. <https://kaggle.com/competitions/diabetic-retinopathy-detection>, 2015. Kaggle.
- Bradley J. Erickson, David Mutch, Lynne Lippmann, and Rose Jarosz. The cancer genome atlas uterine corpus endometrial carcinoma collection (tcga-ucec), 2016. URL <https://www.cancerimagingarchive.net/collection/tcga-ucec/>.
- Enid M Eslick, John Kipritidis, Denis Gradinscak, Mark J Stevens, Dale L Bailey, Benjamin Harris, Jeremy T Booth, and Paul J Keall. Ct ventilation imaging derived from breath hold ct exhibits good regional accuracy with galligas pet. *Radiotherapy and Oncology*, 127(2):267–273, 2018.
- Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pp. 12873–12883, 2021.
- S. Graham et al. CoNIC: Colon Nuclei Identification and Counting Challenge 2022. *arXiv preprint arXiv:2111.14485*, 2021.
- Huihui Fang, Fei Li, Junde Wu, Huazhu Fu, Xu Sun, Jaemin Son, Shuang Yu, Menglu Zhang, Chenglang Yuan, Cheng Bian, et al. Refuge2 challenge: A treasure trove for multi-dimension analysis and evaluation in glaucoma screening. *arXiv preprint arXiv:2202.08994*, 2022.
- Huihui Fang, Fei Li, Junde Wu, Huazhu Fu, Xu Sun, José Ignacio Orlando, Hrvoje Bogunović, Xiulan Zhang, and Yanwu Xu. Open fundus photograph dataset with pathologic myopia recognition and anatomical structure annotation. *Scientific Data*, 11(1):99, 2024.
- Andrey Fedorov, William JR Longabaugh, David Pot, David A Clunie, Steven D Pieper, David L Gibbs, Christopher Bridge, Markus D Herrmann, André Homeyer, Rob Lewis, et al. National cancer institute imaging data commons: toward transparency, reproducibility, and scalability in imaging artificial intelligence. *Radiographics*, 43(12):e230180, 2023.
- Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 59(9):2538–2548, 2012.
- 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*, pp. 11–19, 2019.
- Jevgenij Gamper, Navid Alemi Koohbanani, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. Pannuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778*, 2020.
- Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberger, 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.
- Elisa Drelie Gelasca, Jiyun Byun, Boguslaw Obara, and B.S. Manjunath. Evaluation and benchmark for biological image segmentation. In *IEEE International Conference on Image Processing*, Oct 2008.

- Ioannis Giotis, Nynke Molders, Sander Land, Michael Biehl, Marcel F Jonkman, and Nicolai Petkov. MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Systems with Applications*, 42(19):6578–6585, 2015.
- 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.
- Grace Maria Binu. Retinal occlusion dataset. <https://www.kaggle.com/datasets/gracemariabinu/retinal-occlusion-dataset>, 2023. Kaggle dataset (uploader: Grace Maria Binu). Accessed 2025-08-21.
- Simon Graham, Hao Chen, Jevgenij Gamper, Qi Dou, Pheng-Ann Heng, David Snead, Yee Wah Tsang, and Nasir Rajpoot. Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. *Medical Image Analysis*, 52:199–211, 2019a.
- Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019b.
- 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)*, pp. 1635–1638, 2019.
- 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*, pp. 1820–1828, 2021.
- Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–26, 2022.
- 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.
- 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. Anderson cancer center head and neck quantitative imaging working group. (2020) hnscc version 4 [dataset], 2020. URL <https://www.cancerimagingarchive.net/collection/hnscc>.
- O. Grove, A. E. Berglund, M. B. Schabath, H. J. W. L. Aerts, A. Dekker, H. Wang, E. R. Velazquez, P. Lambin, Y. Gu, Y. Balagurunathan, E. Eikman, R. A. Gatenby, S. Eschrich, and R. J. Gillies. Data from: Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma, 2015. URL <https://doi.org/10.7937/K9/TCIA.2015.A6V7JIWX>. All images are diagnostic contrast-enhanced CT scans obtained at diagnosis, prior to surgery. Study aimed to extract prognostic image features for lung adenocarcinoma.
- Anubha Gupta, Shiv Gehlot, Shubham Goswami, Sachin Motwani, Ritu Gupta, Álvaro García Faura, Dejan Štepec, Tomaž Martinčič, Reza Azad, Dorit Merhof, et al. Segpc-2021: A challenge & dataset on segmentation of multiple myeloma plasma cells from microscopic images. *Medical Image Analysis*, 83:102677, 2023.

- David Gutman, N Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection. In *International Symposium on Biomedical Imaging (ISBI)*, (International Skin Imaging Collaboration (ISIC), 2016), 2016.
- Safwan S Halabi, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Artem B Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, et al. The RSNA pediatric bone age machine learning challenge. *Radiology*, 290(2):498–503, 2019.
- Chu Han, Xipeng Pan, Lixu Yan, Huan Lin, Bingbing Li, Su Yao, Shanshan Lv, Zhenwei Shi, Jinhai Mai, Jiatai Lin, Bingchao Zhao, Zeyan Xu, Zhizhen Wang, Yumeng Wang, Yuan Zhang, Huihui Wang, Chao Zhu, Chunhui Lin, Lijian Mao, Min Wu, Luwen Duan, Jingsong Zhu, Dong Hu, Zijie Fang, Yang Chen, Yongbing Zhang, Yi Li, Yiwen Zou, Yiduo Yu, Xiaomeng Li, Haiming Li, Yanfen Cui, Guoqiang Han, Yan Xu, Jun Xu, Huihua Yang, Chunming Li, Zhenbing Liu, Cheng Lu, Xin Chen, Changhong Liang, Qingling Zhang, and Zaiyi Liu. Wsss4luad: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. *arXiv preprint arXiv:2204.06455*, 2022.
- 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.
- Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, et al. Comparison and evaluation of methods for liver segmentation from ct datasets. *IEEE Transactions on Medical Imaging*, 28(8):1251–1265, 2009.
- Nicholas Heller, Niranjana Sathianathan, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, Joshua Dean, Michael Tradewell, Aneri Shah, Resha Tejpal, Zachary Edgerton, Matthew Peterson, Shaneabbas Raza, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. C4kc kits challenge kidney tumor segmentation dataset, 2019. URL <https://www.cancerimagingarchive.net/collection/c4kc-kits/>.
- Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, pp. 101821, 2020.
- 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, 2022a.
- 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, 2022b.
- Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific Data*, 9(1):762, 2022.
- 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.
- Laurens Hogeweg, Clara I. Sánchez, Pim A. de Jong, Pragnya Maduskar, and Bram van Ginneken. Clavicle segmentation in chest radiographs. *Medical Image Analysis*, 16(8):1490–1502, 2012.

- Chandra Holback, Rose Jarosz, Fred Prior, David G. Mutch, Priya Bhosale, Kimberly Garcia, Yueh Lee, Shanah Kirk, Cheryl A. Sadow, Seth Levine, Evis Sala, Pierre Elnajjar, Tara Morgan, and Bradley J. Erickson. The cancer genome atlas ovarian cancer collection (tcga-ov), 2016. URL <https://www.cancerimagingarchive.net/collection/tcga-ov/>.
- Sven Holm, Greg Russell, Vincent Nourrit, and Niall McLoughlin. Dr hags—a fundus image database for the automatic extraction of retinal surface vessels from diabetic patients. *Journal of Medical Imaging*, 4(1):014503–014503, 2017.
- A.D. Hoover, V. Kouznetsova, and M. Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging*, 19(3):203–210, 2000.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Qiao Hu, Michael D Abràmoff, and Mona K Garvin. Automated separation of binary overlapping trees in low-contrast color retinal images. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22–26, 2013, Proceedings, Part II 16*, pp. 436–443, 2013.
- Arnaud Huaulmé, Kanako Harada, Quang-Minh Nguyen, Bogyu Park, Seungbum Hong, Min-Kook Choi, Michael Peven, Yunshuang Li, Yonghao Long, Qi Dou, et al. PEg TRAnsfer workflow recognition challenge report: Do multimodal data improve recognition? *Computer Methods and Programs in Biomedicine*, 236:107561, 2023.
- Mir Tanvir Islam, Shafin T Mashfu, Abrar Faisal, Sadman Chowdhury Siam, Intisar Tahmid Naheen, and Riasat Khan. Deep learning-based glaucoma detection with cropped optic cup and disc and blood vessel segmentation. *Ieee Access*, 10:2828–2841, 2021.
- Mir Tanvir Islam, Shafin T. Mashfu, Abrar Faisal, Sadman Chowdhury Siam, Intisar Tahmid Naheen, and Riasat Khan. Deep learning-based glaucoma detection with cropped optic cup and disc and blood vessel segmentation. *IEEE Access*, 10:2828–2841, 2022.
- IXI. IXI Dataset. <https://brain-development.org/ixi-dataset/>, 2010. Imperial College London, Creative Commons CC BY-SA 3.0.
- Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K. Singh, Sameer Antani, George Thoma, Yi-Xiang Wang, Pu-Xuan Lu, and Clement J. McDonald. Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging*, 33(2):233–245, 2014a.
- Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K. Singh, Sameer Antani, George Thoma, Yi-Xiang Wang, Pu-Xuan Lu, and Clement J. McDonald. Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging*, 33(2):233–245, 2014b. doi: 10.1109/TMI.2013.2284099.
- Jan Jantzen, Jonas Norup, Georgios Dounias, and Beth Bjerregaard. Pap-smear benchmark data for pattern classification. *Nature inspired smart information systems (NiSIS 2005)*, pp. 1–9, 2005.
- 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.
- Yiping Jiao, Jeroen van der Laak, Shadi Albarqouni, Zhang Li, Tao Tan, Abhir Bhalerao, Shenghua Cheng, Jiabo Ma, Johnathan Pocock, Josien P.W. Pluim, Navid Alemi Koohbanani, Raja Muhammad Saad Bashir, Shan E Ahmed Raza, Sibo Liu, Simon Graham, Suzanne Wetstein, Syed Ali Khurram, Xiuli Liu, Nasir Rajpoot, Mitko Veta, and Francesco Ciompi. LYSTO: The lymphocyte assessment hackathon and benchmark dataset. *IEEE Journal of Biomedical and Health Informatics*, 28(3):1161–1172, 2024.



- Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. *IEEE Winter Conference on Applications of Computer Vision*, 2018.
- Kai Jin, Xingru Huang, Jingxing Zhou, Yunxiang Li, Yan Yan, Yibao Sun, Qianni Zhang, Yaqi Wang, and Juan Ye. FIVES: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific Data*, 9(1):475, 2022.
- 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.
- 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.
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pp. 694–711, 2016.
- 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/pelvic ct exams with expert organ contours (pediatric-ct-seg) (version 2) [data set], 2021. URL <https://www.cancerimagingarchive.net/collection/pediatric-ct-seg>.
- 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.
- Ma Jun, Ge Cheng, Wang Yixin, An Xingle, Gao Jiantao, Yu Ziqi, Zhang Mingqing, Liu Xin, Deng Xueyuan, Cao Shucheng, Wei Hao, Mei Sen, Yang Xiaoyu, Nie Ziwei, Li Chen, Tian Lu, Zhu Yuntao, Zhu Qiongjie, Dong Guoqiang, and He Jian. Covid-19 ct lung and infection segmentation dataset, April 2020. URL <https://doi.org/10.5281/zenodo.3757476>.
- J. Kalpathy-Cramer, S. Napel, D. Goldgof, and B. Zhao. Multi-site collection of lung ct data with nodule segmentations (version 3), 2015. URL <https://doi.org/10.7937/k9/tcia.2015.1buvfjr7>. CT images and segmentations of non-small cell lung cancer from multiple institutions, designed for algorithm development and comparison.
- Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Medicine*, 16(1):e1002730, 2019.
- A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debodoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.
- Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, pp. 538–546, 2018.
- Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Debanjan Haldar, Zhifan Jiang, Syed Muhammed Anwar, Jake Albrecht, Maruf Adewole, Udunna Anazodo, Hannah Anderson, Sina Bagheri, Ujjwal Baid, Timothy Bergquist, Austin J. Borja, Evan Calabrese, Verena

- Chung, Gian-Marco Conte, Farouk Dako, James Eddy, Ivan Ezhov, Ariana Familiar, Keyvan Farahani, Shuvanjan Haldar, Juan Eugenio Iglesias, Anastasia Janas, Elaine Johansen, Blaise V Jones, Florian Kofler, Dominic LaBella, Hollie Anne Lai, Koen Van Leemput, Hongwei Bran Li, Nazanin Maleki, Aaron S McAllister, Zeke Meier, Bjoern Menze, Ahmed W Moawad, Khanak K Nandolia, Julija Pavaine, Marie Piraud, Tina Poussaint, Sanjay P Prabhu, Zachary Reitman, Andres Rodriguez, Jeffrey D Rudie, Ibraheem Salman Shaikh, Lubdha M. Shah, Nakul Sheth, Russel Taki Shinohara, Wenxin Tu, Karthik Viswanathan, Chunhao Wang, Jeffrey B Ware, Benedikt Wiestler, Walter Wiggins, Anna Zapaishchykova, Mariam Aboian, Miriam Bornhorst, Peter de Blank, Michelle Deutsch, Maryam Fouladi, Lindsey Hoffman, Benjamin Kann, Margot Lazow, Leonie Mikael, Ali Nabavizadeh, Roger Packer, Adam Resnick, Brian Rood, Arastoo Vossough, Spyridon Bakas, and Marius George Linguraru. The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtbn-connect-dipgr-asnr-miccai brats-peds), 2023.
- Daniel S Kermay, 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.
- Ungsoo Kim. Machine learn for glaucoma, 2018. URL <https://doi.org/10.7910/DVN/1YRRAC>. Accessed: 2025-05-22.
- 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>.
- P. Kinahan, M. Muzi, B. Bialecki, B. Herman, and L. Coombs. Data from the acrin 6668 trial nsccl-fdg-pet (version 2) [data set], 2019. URL <https://www.cancerimagingarchive.net/collection/acrin-nsccl-fdg-pet>.
- Paul Kinahan, Mark Muzi, Brian Bialecki, and Laura Coombs. Data from the acrin 6685 trial hnscc-fdg-pet/ct, 2020. URL <https://www.cancerimagingarchive.net/collection/acrin-hnscc-fdg-pet-ct/>.
- Diederik P Kingma. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014. URL <https://arxiv.org/abs/1412.6980>.
- Serkan Kiranyaz, Aysen Degerli, Tahir Hamid, Rashid Mazhar, Rayyan El Fadil Ahmed, Rayaana Abouhasera, Morteza Zabihi, Junaid Malik, Ridha Hamila, and Moncef Gabbouj. Left ventricular wall motion estimation by active polynomials for acute myocardial infarction detection. *IEEE Access*, 8:210301–210317, 2020.
- S. Kirk, Y. Lee, C. A. Sadow, and S. Levine. The cancer genome atlas rectum adenocarcinoma collection (tcga-read) (version 3) [data set], 2016a. URL <https://www.cancerimagingarchive.net/collection/tcga-read>.
- Shanahv Kirk, Yueh Lee, Prasanna Kumar, Joe Filippini, Brad Albertina, Mark Watson, Kimberly Rieger-Christ, and John Lemmerman. The cancer genome atlas lung squamous cell carcinoma collection (tcga-lusc), 2016b. URL <https://www.cancerimagingarchive.net/collection/tcga-lusc/>.
- Arno Klein, Jesper Andersson, Babak A. Ardekani, John Ashburner, Brian Avants, Ming-Chang Chiang, Gary E. Christensen, D. Louis Collins, James Gee, Pierre Hellier, Joo Hyun Song, Mark Jenkinson, Claude Lepage, Daniel Rueckert, Paul Thompson, Tom Vercauteren, Roger P. Woods, J. John Mann, and Ramin V. Parsey. Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. *NeuroImage*, 46(3):786–802, 2009.
- Arno Klein, Satrajit S. Ghosh, Forrest S. Bao, Joachim Giard, Yrjö Häme, Eliezer Stavsky, Noah Lee, Brian Rossa, Martin Reuter, Elias Chaibub Neto, and Anisha Keshavan. Mindboggling morphometry of human brains. *PLOS Computational Biology*, 13(2):1–40, 02 2017.
- Oleksandr Kovalyk, Juan Morales-Sánchez, Rafael Verdú-Monedero, Inmaculada Sellés-Navarro, Ana Palazón-Cabanes, and José-Luis Sancho-Gómez. Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Scientific Data*, 9(1):291, 2022.

- Hugo J. Kuijff, J. Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M. Jorge Cardoso, Adrià Casamitjana, D. Louis Collins, Mahsa Dadar, Achilleas Georgiou, Mohsen Ghafoorian, Dakai Jin, April Khademi, Jesse Knight, Hongwei Li, Xavier Lladó, Miguel Luna, Qaiser Mahmood, Richard McKinley, Alireza Mehrtaash, Sébastien Ourselin, Bo-Yong Park, Hyunjin Park, Sang Hyun Park, Simon Pezold, Elodie Puybareau, Leticia Rittner, Carole H. Sudre, Sergi Valverde, Verónica Vilaplana, Roland Wiest, Yongchao Xu, Ziyue Xu, Guodong Zeng, Jianguo Zhang, Guoyan Zheng, Christopher Chen, Wiesje van der Flier, Frederik Barkhof, Max A. Viergever, and Geert Jan Biessels. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE Transactions on Medical Imaging*, 38(11):2556–2568, 2019.
- Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, Yunzhi Wang, Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajeddin, Ali Gooya, Nasir Rajpoot, Xuhua Ren, Sihang Zhou, Qian Wang, Dinggang Shen, Cheng-Kun Yang, Chi-Hung Weng, Wei-Hsiang Yu, Chao-Yuan Yeh, Shuang Yang, Shuoyu Xu, Pak Hei Yeung, Peng Sun, Amirreza Mahbod, Gerald Schaefer, Isabella Ellinger, Rupert Ecker, Orjan Smedby, Chunliang Wang, Benjamin Chidester, That-Vinh Ton, Minh-Triet Tran, Jian Ma, Minh N. Do, Simon Graham, Quoc Dang Vu, Jin Tae Kwak, Akshaykumar Gunda, Raviteja Chunduri, Corey Hu, Xiaoyang Zhou, Dariush Lotfi, Reza Safdari, Antanas Kascenas, Alison O’Neil, Dennis Eschweiler, Johannes Stegmaier, Yanping Cui, Baocai Yin, Kailin Chen, Xinmei Tian, Philipp Gruening, Erhardt Barth, Elad Arbel, Itay Remer, Amir Ben-Dor, Ekaterina Sirazitdinova, Matthias Kohl, Stefan Braunewell, Yuexiang Li, Xinpeng Xie, Linlin Shen, Jun Ma, Krishanu Das Baksi, Mohammad Azam Khan, Jaegul Choo, Adrián Colomer, Valery Naranjo, Linmin Pei, Khan M. Iftexharuddin, Kaushiki Roy, Debotosh Bhattacharjee, Anibal Pedraza, Maria Gloria Bueno, Sabarinathan Devanathan, Saravanan Radhakrishnan, Praveen Koduganty, Zihan Wu, Guanyu Cai, Xiaojie Liu, Yuqin Wang, and Amit Sethi. A multi-organ nucleus segmentation challenge. *IEEE Transactions on Medical Imaging*, 39(5):1380–1391, 2020.
- K.A. et al. Kurdziel. Data from NaF PROSTATE. The Cancer Imaging Archive Dataset, 2015.
- Dominic LaBella, Maruf Adewole, Michelle Alonso-Basanta, Talissa Altes, Syed Muhammad Anwar, Ujjwal Baid, Timothy Bergquist, Radhika Bhalariao, Sully Chen, Verena Chung, et al. The asnr-miccai brain tumor segmentation (brats) challenge 2023: Intracranial meningioma. *arXiv preprint arXiv:2305.07642*, 2023.
- Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: Segmentation of thoracic organs at risk in ct images. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6, 2020.
- Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. MICCAI multi-atlas labeling beyond the cranial vault—workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, pp. 12, 2015.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):1–10, 2018.
- Patrick Leavey, Anita Sengupta, Dinesh Rakheja, Ovidiu Daescu, Harish B Arunachalam, and Rashika Mishra. Osteosarcoma data from ut southwestern/ut dallas for viable and necrotic tumor assessment [data set]. *Cancer Imaging Arch*, 14, 2019.
- 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. *arXiv:2306.00890*, 2023.
- Liu Li, Mai Xu, Hanruo Liu, Yang Li, Xiaofei Wang, Lai Jiang, Zulin Wang, Xiang Fan, and Ningli Wang. A large-scale database and a cnn model for attention-based glaucoma detection. *IEEE Transactions on Medical Imaging*, 39(2):413–424, 2019.
- 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>.

- Tianbin Li, Yanzhou Su, Wei Li, Bin Fu, Zhe Chen, Ziyang Huang, Guoan Wang, Chenglong Ma, Ying Chen, Ming Hu, Yanjun Li, Pengcheng Chen, Xiaowei Hu, Zhongying Deng, Yuanfeng Ji, Jin Ye, Yu Qiao, and Junjun He. GMAI-VL & GMAI-VL-5.5M: A large vision-language model and a comprehensive multimodal dataset towards general medical AI. *arXiv preprint arXiv:2411.14522*, 2024.
- Xia Li, Richard G Abramson, Lori R Arlinghaus, Hakmook Kang, Anuradha Bapsi Chakravarthy, Vandana G Abramson, Jaime Farley, Ingrid A Mayer, Mark C Kelley, Ingrid M Meszoely, et al. Multiparametric magnetic resonance imaging for predicting pathological response after the first cycle of neoadjuvant chemotherapy in breast cancer. *Investigative Radiology*, 50(4):195–204, 2015.
- Jie Lian, Jingyu Liu, Shu Zhang, Kai Gao, Xiaoqing Liu, Dingwen Zhang, and Yizhou Yu. A structure-aware relation network for thoracic diseases detection and segmentation. *IEEE Transactions on Medical Imaging*, 40(8):2042–2052, 2021.
- 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.
- Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wangui He, Hao Jiang, Mengze Li, Xiaohui Song, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Chin Beng Ooi. HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.
- Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-CLIP: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 525–536, 2023.
- M. W. Linehan, R. Gautam, C. A. Sadow, and S. Levine. The cancer genome atlas kidney chromophobe collection (tcga-kich), 2016. URL <https://doi.org/10.7937/K9/TCIA.2016.YU3RBCZN>. Data set.
- 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*, pp. 1650–1654, 2021a.
- Chi Liu, Xiaotong Han, Zhixi Li, Jason Ha, Guankai Peng, Wei Meng, and Mingguang He. A self-adaptive deep learning method for automated eye laterality detection based on color fundus photography. *PloS One*, 14(9):e0222025, 2019.
- 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, 2021b.
- Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, et al. DeepDRiD: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3(6), 2022.
- 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.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- 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>.

- Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Austin Wolfgang Katzer, Collin Chiu, et al. BIOMEDICA: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. *arXiv preprint arXiv:2501.07171*, 2025.
- Valeria De Luca, Tobias Benz, Satoshi Kondo, Christine Tanner, and Orcun Goksel. Evaluation of 2d and 3d ultrasound tracking algorithms and impact on ultrasound-guided liver radiotherapy margins. *Medical Physics*, 45(11):5105–5116, 2018. doi: 10.1002/mp.13147. URL <https://doi.org/10.1002/mp.13147>.
- Fabiano R. Lucchesi and Natália D. Aredes. The cancer genome atlas stomach adenocarcinoma collection (tcga-stad), 2016. URL <https://www.cancerimagingarchive.net/collection/tcga-stad/>.
- 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*, 2023a.
- 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*, 2023b.
- Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N Metaxas, Guotai Wang, and Shaoting Zhang. WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642, 2022.
- Chenglong Ma, Yuanfeng Ji, Jin Ye, Lu Zhang, Ying Chen, Tianbin Li, Mingjie Li, Junjun He, and Hongming Shan. Towards interpretable counterfactual generation via multimodal autoregression. *arXiv preprint arXiv:2503.23149*, 2025a.
- 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. *arXiv:2502.20321*, 2025b.
- 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, 2021a.
- Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Mae, Adamo Young, Cheng Zhu, Xin Yang, Kangkang Meng, Ziyang 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.
- Yuhui Ma, Huaying Hao, Jianyang Xie, Huazhu Fu, Jiong Zhang, Jianlong Yang, Zhen Wang, Jiang Liu, Yalin Zheng, and Yitian Zhao. Rose: a retinal oct-angiography vessel segmentation dataset and new model. *IEEE Transactions on Medical Imaging*, 40(3):928–939, 2021b.
- 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.
- 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](https://www.cancerimagingarchive.net/collection/anti-pd-1_lung).
- 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.



- Mojtaba Masoudi, Hamid-Reza Pourreza, Mahdi Saadatmand-Tarzjan, Noushin Eftekhari, Fateme Shafiee Zargar, and Masoud Pezeshki Rad. A new dataset of computed-tomography angiography images for computer-aided detection of pulmonary embolism. *Scientific Data*, 5(1): 1–9, 2018.
- 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.
- Cynthia H McCollough, Adam C Bartley, Rickey E Carter, Baiyu Chen, Tammy A Drees, Phillip Edwards, David R Holmes III, Alice E Huang, Farhana Khan, Shuai Leng, et al. Low-dose CT for the detection and classification of metastatic liver lesions: results of the 2016 low dose CT grand challenge. *Medical Physics*, 44(10):e339–e352, 2017.
- 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.
- Teresa Mendonça, M Celebi, T Mendonca, and J Marques. Ph2: A public database for the analysis of dermoscopic images. *Dermoscopy Image Analysis*, 2, 2015.
- Ahmed W. Moawad, Anastasia Janas, Ujjwal Baid, Divya Ramakrishnan, Leon Jekel, Kiril Krantchev, Harrison Moy, Rachit Saluja, Klara Osenberg, Klara Wilms, Manpreet Kaur, Arman Avesta, Gabriel Cassinelli Pedersen, Nazanin Maleki, Mahdi Salimi, Sarah Merkaj, Marc von Reppert, Niklas Tillmans, Jan Lost, Khaled Bousabarah, Wolfgang Holler, MingDe Lin, Malte Westerhoff, Ryan Maresca, Katherine E. Link, Nourel hoda Tahon, Daniel Marcus, Aristeidis Sotiras, Pamela LaMontagne, Strajit Chakrabarty, Oleg Teytelboym, Ayda Youssef, Ayaman Nada, Yuri S. Velichko, Nicolo Gennaro, Connectome Students, Group of Annotators, Justin Cramer, Derek R. Johnson, Benjamin Y. M. Kwan, Boyan Petrovic, Satya N. Patro, Lei Wu, Tiffany So, Gerry Thompson, Anthony Kam, Gloria Guzman Perez-Carrillo, Neil Lall, Group of Approvers, Jake Albrecht, Udunna Anazodo, Marius George Lingaru, Bjoern H Menze, Benedikt Wiestler, Maruf Adewole, Syed Muhammad Anwar, Dominic Labella, Hongwei Bran Li, Juan Eugenio Iglesias, Keyvan Farahani, James Eddy, Timothy Bergquist, Verena Chung, Russel Takeshi Shinohara, Farouk Dako, Walter Wiggins, Zachary Reitman, Chunhao Wang, Xinyang Liu, Zhifan Jiang, Koen Van Leemput, Marie Piraud, Ivan Ezhov, Elaine Johanson, Zeke Meier, Ariana Familiar, Anahita Fathi Kazerooni, Florian Kofler, Evan Calabrese, Sanjay Aneja, Veronica Chiang, Ichiro Ikuta, Umber Shafique, Fatima Memon, Gian Marco Conte, Spyridon Bakas, Jeffrey Rudie, and Mariam Aboian. The brain tumor segmentation (BraTS-METS) challenge 2023: Brain metastasis segmentation on pre-treatment mri. *arXiv preprint arXiv:2306.00838*, 2023.
- Anna Montoya, Hasnin, kaggle446, shirzad, Will Cukierski, and yffud. Ultrasound nerve segmentation. <https://kaggle.com/competitions/ultrasound-nerve-segmentation>, 2016. Kaggle.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225, pp. 353–367, 2023.
- Simmi Mourya, Sonaal Kant, Pulkit Kumar, Anubha Gupta, and Rita Gupta. All challenge dataset of isbi 2019 (c-nmc 2019), 2019. URL <https://www.cancerimagingarchive.net/collection/c-nmc-2019/>.
- 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.

- National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium clear cell renal cell carcinoma collection (cptac-ccrcc), 2018a. URL <https://www.cancerimagingarchive.net/collection/cptac-ccrcc/>.
- National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium cutaneous melanoma collection (cptac-cm), 2018b. URL <https://www.cancerimagingarchive.net/collection/cptac-cm/>.
- 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>.
- 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>.
- National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium uterine corpus endometrial carcinoma collection (cptac-ucec), 2019a. URL <https://www.cancerimagingarchive.net/collection/cptac-ucec/>.
- National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium sarcomas collection (cptac-sar), 2019b. URL <https://www.cancerimagingarchive.net/collection/cptac-sar/>.
- Ilya Nelkenbaum, Galia Tsarfaty, Nahum Kiryati, Eli Konen, and Arnaldo Mayer. Automatic segmentation of white matter tracts using multiple brain mri sequences. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 368–371, 2020.
- Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330, 2012.
- Duc Nguyen, DungNB, Ha Q. Nguyen, Julia Elliott, NguyenThanhNhan, and Phil Culliton. Vinbigdata chest x-ray abnormalities detection. <https://kaggle.com/competitions/vinbigdata-chest-xray-abnormalities-detection>, 2020. Kaggle.
- Uyen TV Nguyen, Alauddin Bhuiyan, Laurence AF Park, Ryo Kawasaki, Tien Y Wong, Jie Jin Wang, Paul Mitchell, and Kotagiri Ramamohanarao. An automated method for retinal arteriovenous nicking quantification from color fundus images. *IEEE Transactions on Biomedical Engineering*, 60(11):3194–3203, 2013.
- Guy Nir, Soheil Hor, Davood Karimi, Ladan Fazli, Brian F Skinnider, Peyman Tavassoli, Dmitry Turbin, Carlos F Villamil, Gang Wang, R Storey Wilson, et al. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical Image Analysis*, 50: 167–180, 2018a.
- Guy Nir, Soheil Hor, Davood Karimi, Ladan Fazli, Brian F Skinnider, Peyman Tavassoli, Dmitry Turbin, Carlos F Villamil, Gang Wang, R Storey Wilson, et al. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical Image Analysis*, 50: 167–180, 2018b.
- Chinedu Innocent Nwoye, Deepak Alapatt, Tong Yu, Armine Vardazaryan, Fangfang Xia, Zixuan Zhao, Tong Xia, Fucang Jia, Yuxuan Yang, Hao Wang, Derong Yu, Guoyan Zheng, Xiaotian Duan, Neil Getty, Ricardo Sanchez-Matilla, Maria Robu, Li Zhang, Huabin Chen, Jiacheng Wang, Liansheng Wang, Bokai Zhang, Beerend Gerats, Sista Raviteja, Rachana Sathish, Rong Tao, Satoshi Kondo, Winnie Pang, Hongliang Ren, Julian Ronald Abbing, Mohammad Hasan Sarhan, Sebastian Bodenstedt, Nithya Bhasker, Bruno Oliveira, Helena R. Torres, Li Ling, Finn Gaida, Tobias Czempel, João L. Vilça, Pedro Morais, Jaime Fonseca, Ruby Mae Egging, Inge Nicole Wijma, Chen Qian, Guibin Bian, Zhen Li, Velmurugan Balasubramanian, Debdoot Sheet, Imanol Luengo, Yuanbo Zhu, Shuai Ding, Jakob-Anton Aschenbrenner, Nicolas Elini van der Kar, Mengya

- Xu, Mobarakol Islam, Lalithkumar Seenivasan, Alexander Jenke, Danail Stoyanov, Didier Mutter, Pietro Mascagni, Barbara Seeliger, Cristians Gonzalez, and Nicolas Padoy. Cholectriple2021: A benchmark challenge for surgical action triplet recognition. *Medical Image Analysis*, 86:102803, 2023.
- OpenAI. Addendum to GPT-4o system card: Native image generation. Technical report, OpenAI, March 2025. URL [https://cdn.openai.com/11998be9-5319-4302-bfbf-1167e093f1fb/Native\\_Image\\_Generation\\_System\\_Card.pdf](https://cdn.openai.com/11998be9-5319-4302-bfbf-1167e093f1fb/Native_Image_Generation_System_Card.pdf). Accessed: 2025-05-16.
- OpenMEDLab. MedFM ChestDR 2023 Chest X-Ray Disease Screening Dataset. NeurIPS MedFM Challenge Dataset, 2023.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- StructSeg Challenge Organizers. StructSeg2019: Head & Neck OAR Segmentation. MICCAI 2019 Challenge Dataset, 2019.
- Silvia Ovreiu, Elena-Anca Paraschiv, and Elena Ovreiu. Deep learning & digital fundus images: Glaucoma detection using densenet. In *2021 13th International Conference on Electronics, Computers and Artificial Intelligence*, pp. 1–4, 2021.
- Danielle F Pace, Adrian V Dalca, Tal Geva, Andrew J Powell, Mehdi H Moghari, and Polina Golland. Interactive whole-heart segmentation in congenital heart disease. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 80–88. Springer, 2015.
- Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:106221, 2020.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.
- Sarthak Pati, Ruchika Verma, Hamed Akbari, Michel Bilello, Virginia B Hill, Chiharu Sako, Ramon Correa, Niha Beig, Ludovic Venet, Siddhesh Thakur, et al. Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the ivy glioblastoma atlas project (Ivy GAP) dataset. *Medical Physics*, 47(12):6039–6052, 2020.
- M. Patnana, S. Patel, and A. Tsao. Anti-PD-1 immunotherapy melanoma dataset [data set], 2019.
- Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *10th International Symposium on Medical Information Processing and Analysis*, volume 9287, pp. 188–193, 2015.
- João Pedrosa, Guilherme Aresta, Carlos Ferreira, Márcio Rodrigues, Patrícia Leitão, André Silva Carvalho, João Rebelo, Eduardo Negrão, Isabel Ramos, António Cunha, et al. LNDdb: a lung nodule database on computed tomography. *arXiv preprint arXiv:1911.08434*, 2019.
- Peking University International Competition on Ocular Disease Intelligent Recognition (ODIR-2019). Ocular disease intelligent recognition (odir-2019) [data set], 2019. URL <https://odir2019.grand-challenge.org/introduction/>. Accessed: 2025-05-22.
- Yifan Peng, Yuxing Tang, Sungwon Lee, Yingying Zhu, Ronald M Summers, and Zhiyong Lu. Covid-19-ct-cxr: a freely accessible and weakly labeled chest x-ray and ct image collection on covid-19 from biomedical literature. *IEEE transactions on big data*, 7(1):3–12, 2020.

- Etta D Pisano and Martin J Yaffe. Digital mammography. *Radiology*, 234(2):353–362, 2005.
- Marco Pizzolato, Marco Palombo, Elisenda Bonet-Carne, Chantal M. W. Tax, Francesco Grussu, Andrada Ianus, Fabian Bogusz, Tomasz Pieciak, Lipeng Ning, Hugo Larochelle, Maxime Descoteaux, Maxime Chamberland, Stefano B. Blumberg, Thomy Mertzanidou, Daniel C. Alexander, Maryam Afzali, Santiago Aja-Fernández, Derek K. Jones, Carl-Fredrik Westin, Yogesh Rath, Steven H. Baete, Lucilio Cordero-Grande, Thilo Ladner, Paddy J. Sator, Joseph V. Hajnal, Jean-Philippe Thiran, Anthony N. Price, Farshid Sepehrband, Fan Zhang, and Jana Hutter. Acquiring and predicting multidimensional diffusion (mudi) data: An open challenge. In Elisenda Bonet-Carne, Jana Hutter, Marco Palombo, Marco Pizzolato, Farshid Sepehrband, and Fan Zhang (eds.), *Computational Diffusion MRI*, pp. 195–208, 2020.
- Alibaba Tianchi Platform. Diabetic retinopathy arranged dataset. <https://tianchi.aliyun.com/dataset/93926>, 2023. Accessed: 2021-03-10.
- Gašper Podobnik, Primož Strojani, Primož Peterlin, Bulat Ibragimov, and Tomaž Vrtovec. Hansseg: The head and neck organ-at-risk ct and mr segmentation dataset. *Medical Physics*, 50(3):1917–1927, 2023.
- Prasanna Porwal, Samiksha Pachade, Manesh Kokare, Girish Deshmukh, Jaemin Son, Woong Bae, Lihong Liu, Jianzong Wang, Xinhui Liu, Liangxin Gao, TianBo Wu, Jing Xiao, Fengyan Wang, Baocai Yin, Yunzhi Wang, Gopichandh Danala, Linsheng He, Yoon Ho Choi, Yeong Chan Lee, Sang-Hyuk Jung, Zhongyu Li, Xiaodan Sui, Junyan Wu, Xiaolong Li, Ting Zhou, Janos Toth, Agnes Baran, Avinash Kori, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Xingzheng Lyu, Li Cheng, Qin hao Chu, Pengcheng Li, Xin Ji, Sanyuan Zhang, Yaxin Shen, Ling Dai, Oindrila Saha, Rachana Sathish, Tânia Melo, Teresa Araújo, Balazs Harangi, Bin Sheng, Ruogu Fang, Debodoot Sheet, Andras Hajdu, Yuanjie Zheng, Ana Maria Mendonça, Shaoting Zhang, Aurélio Campilho, Bin Zheng, Dinggang Shen, Luca Giancardo, Gwennolé Quéllec, and Fabrice Mériaudeau. IDRiD: Diabetic retinopathy – segmentation and grading challenge. *Medical Image Analysis*, 59:101561, 2020.
- Mohit Prabhushankar, Kiran Kokilepersaud, Yash-ye 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.
- Praveen Govi. Coronahack - chest x-ray-dataset. <https://www.kaggle.com/datasets/praveengovi/coronahack-chest-xraydataset>, 2019. Kaggle dataset (uploader: praveengovi). Accessed 2025-08-21.
- Pavle Prentašić, Sven Lončarić, Zoran Vatauvuk, Goran Benčić, Marko Subašić, Tomislav Petković, Lana Dujmović, Maja Malenica-Ravlić, Nikolina Budimlija, and Rašeljka Tadić. Diabetic retinopathy image database (dridb): a new database for diabetic retinopathy screening programs research. In *2013 8th International symposium on Image and Signal Processing and Analysis (ISPA)*, pp. 711–716, 2013.
- Bo Qian, Hao Chen, Xiangning Wang, Haoxuan Che, Gitaek Kwon, Jaeyoung Kim, Sungjin Choi, Seoyoung Shin, Felix Krause, Markus Unterdechler, et al. Drac: diabetic retinopathy analysis challenge with ultra-wide optical coherence tomography angiography images. *arXiv preprint arXiv:2304.02389*, 2023.
- 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. *arXiv preprint arXiv:2412.03069*, 2024.
- Félix Quinton, Romain Popoff, Benoît Presles, Sarah Leclerc, Fabrice Meriaudeau, Guillaume Nodari, Olivier Lopez, Julie Pellegrinelli, Olivier Chevallier, Dominique Gin hac, Jean-Marc Vrigneaud, and Jean-Louis Alberini. A tumour and liver automatic segmentation (atlas) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma. *Data*, 8(5), 2023.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.



- Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. MURA: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*, 2017.
- Sucheng Ren, Xiaoke Huang, Xianhang Li, Junfei Xiao, Jieru Mei, Zeyu Wang, Alan Yuille, and Yuyin Zhou. Medical vision generalist: Unifying medical imaging tasks in context. *arXiv:2406.05565*, 2024.
- 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.
- 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.
- Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I* 18, pp. 556–564. Springer, 2015.
- 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.
- Aurko Roy, Ashish Vaswani, Arvind Neelakantan, and Niki Parmar. Theory and experiments on vector quantized autoencoders. *arXiv preprint arXiv:1805.11063*, 2018.
- Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, et al. ROCOv2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1):688, 2024.
- Rina D. Rudyanto, Sjoerd Kerkstra, Eva M. van Rikxoort, Catalin Fetita, Pierre-Yves Brillet, Christophe Lefevre, Wenzhe Xue, Xiangjun Zhu, Jianming Liang, İlkey Öksüz, Devrim Ünay, Kamuran Kadipaşaoğlu, 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.
- 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>.
- Abdullah Sarhan, Jon Rokne, Reda Alhajj, and Andrew Crichton. Transfer learning through weighted loss function and group normalization for vessel segmentation from retinal images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 9211–9218, 2021.
- Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd Münzer, Manfred Jürgen Primus, and Doris Putzgruber. Cataract-101: video dataset of 101 cataract surgeries. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pp. 421–425, 2018.

- Jarrel Seah, Jen, Maggie, Meng Law, Phil Culliton, and Sarah Dowd. Ranzcr clip - catheter and line position challenge. <https://kaggle.com/competitions/ranzcr-clip-catheter-line-classification>, 2020. Kaggle.
- Anjany Sekuboyina, Malek E Hussein, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, et al. Verse: a vertebrae labelling and segmentation benchmark for multi-detector ct images. *Medical Image Analysis*, 73:102166, 2021a.
- Anjany Sekuboyina, Malek E. Hussein, 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 Huynh, Nicolas Boutry, Roman Jakubicek, Jiri Chmelik, Supriti Mulay, Mohanasankar Sivaprakasam, Johannes C. Paetzold, Suprosanna Shit, Ivan Ezhov, Benedikt Wiestler, Ben Glocker, Alexander Valentinitich, 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, 2021b.
- Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Medical Image Analysis*, 42: 1–13, 2017.
- Ahmed Shahin, Carmela Wegworth, David, Elizabeth Estes, Julia Elliott, Justin Zita, Simon Walsh, Slepety, and Will Cukierski. Osic pulmonary fibrosis progression. <https://kaggle.com/competitions/osic-pulmonary-fibrosis-progression>, 2020. Kaggle.
- 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.
- S. Shikamaru. Glaucoma detection [data set], 2021. URL <https://www.kaggle.com/datasets/sshikamaru/glaucoma-detection>. Accessed: 2025-05-22.
- Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Koda, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.
- Nadya Shusharina and Thomas Bortfeld. Glioma image segmentation for radiotherapy: Rt targets, barriers to cancer spread, and organs at risk (glis-rt), 2021. URL <https://www.cancerimagingarchive.net/collection/glis-rt/>.
- Julio Silva-Rodríguez, Adrián Colomer, María A Sales, Rafael Molina, and Valery Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer Methods and Programs in Biomedicine*, 195: 105637, 2020.
- Korsuk Sirinukunwattana, Josien P.W. Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J. Matuszewski, Elia Bruni, Urko Sanchez, Anton Böhm, Olaf Ronneberger, Bassem Ben Cheikh, Daniel Racoceanu, Philipp Kainz, Michael Pfeiffer, Martin Urschler, David R.J. Snead, and Nasir M. Rajpoot. Gland segmentation in colon histology images: The glas challenge contest. *Medical Image Analysis*, 35:489–502, 2017.

- Jayanthi Sivaswamy, SR Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A Ujjwaft Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pp. 53–56, 2014.
- Ecem Sogancioglu, Bram Van Ginneken, Finn Behrendt, Marcel Bengs, Alexander Schlaefer, Miron Radu, Di Xu, Ke Sheng, Fabien Scalzo, Eric Marcus, et al. Nodule detection and generation on chest x-rays: Node21 challenge. *IEEE Transactions on Medical Imaging*, 2024.
- Md Fahimuzzman Sohan. So you need datasets for your covid-19 detection research using machine learning? *arXiv preprint arXiv:2008.05906*, 2020.
- 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 2022a. URL <https://doi.org/10.5281/zenodo.6575197>.
- 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 2), May 2022b. URL <https://doi.org/10.5281/zenodo.6575270>.
- 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 3), May 2022c. URL <https://doi.org/10.5281/zenodo.6575407>.
- 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 (validation set and annotation), May 2022d. URL <https://doi.org/10.5281/zenodo.6597131>.
- Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2015.
- J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, and B. van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004.
- Sergii Stirenko, Yuriy Kochura, Oleg Alienin, Oleksandr Rokovyi, Yuri Gordienko, Peng Gang, and Wei Zeng. Chest x-ray analysis of tuberculosis by deep learning with segmentation and augmentation. In *2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO)*, pp. 422–428, 2018.
- Martin Styner, Joohwi Lee, Brian Chin, M Chin, Olivier Commowick, H Tran, Silva Markovic-Plese, Valerie Jewells, and Simon Warfield. 3d segmentation in the clinic: A grand challenge ii: Ms lesion segmentation. *MIDAS journal*, 2008:1–6, 2008.
- Sanjay Subramanian, Lucy Lu Wang, Ben Bogin, Sachin Mehta, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. MedICaT: A dataset of medical images, captions, and textual references. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2112–2120, 2020.
- Carole H Sudre, Kimberlin Van Wijnen, Florian Dubost, Hieab Adams, David Atkinson, Frederik Barkhof, Mahlet A Birhanu, Esther E Bron, Robin Camarasa, Nish Chaturvedi, et al. Where is valdo? vascular lesions detection and segmentation challenge at miccai 2021. *Medical Image Analysis*, 91:103029, 2024.
- Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.
- Siham Tabik, Anabel Gómez-Ríos, José Luis Martín-Rodríguez, Iván Sevillano-García, Manuel Rey-Area, David Charte, Emilio Guirado, Juan-Luis Suárez, Julián Luengo, MA Valero-González, et al. COVIDGR dataset and COVID-SDNet methodology for predicting covid-19 based on chest x-ray images. *IEEE Journal of Biomedical and Health Informatics*, 24(12):3595–3605, 2020.

- Ryutaro Tanno, David G. T. Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, Karan Singhal, Mike Schaekermann, Rhys May, Roy Lee, SiWai Man, Sara Mahdavi, Zahra Ahmed, Yossi Matias, Joelle Barral, S. M. Ali Eslami, Danielle Belgrave, Yun Liu, Sreenivasa Raju Kalidindi, Shravya Shetty, Vivek Natarajan, Pushmeet Kohli, Po-Sen Huang, Alan Karthikesalingam, and Ira Ktena. Collaboration between clinicians and vision-language models in radiology report generation. *Nature Medicine*, 31(2):599–608, 2025.
- National Lung Screening Trial Research Team. The national lung screening trial: overview and study design. *Radiology*, 258(1):243–253, 2011.
- Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.
- Kimberley M. Timmins, Irene C. van der Schaaf, Edwin Bennink, Ynte M. Ruigrok, Xingle An, Michael Baumgartner, Pascal Bourdon, Riccardo De Feo, Tommaso Di Noto, Florian Dubost, Augusto Fava-Sanches, Xue Feng, Corentin Giroud, Inteneural Group, Minghui Hu, Paul F. Jaeger, Juhana Kaiponen, Michał Klimont, Yuexiang Li, Hongwei Li, Yi Lin, Timo Loehr, Jun Ma, Klaus H. Maier-Hein, Guillaume Marie, Bjoern Menze, Jonas Richiardi, Saifeddine Rjiba, Dhaval Shah, Suprosanna Shit, Jussi Tohka, Thierry Urruty, Urszula Walińska, Xiaoping Yang, Yunqiao Yang, Yin Yin, Birgitta K. Velthuis, and Hugo J. Kuijf. Comparing methods of detecting and segmenting unruptured intracranial aneurysms on tof-mras: The adam challenge. *NeuroImage*, 238:118216, 2021.
- T Tong and M Li. Abdominal or pelvic enhanced ct images within 10 days before surgery of 230 patients with stage ii colorectal cancer (stageii-colorectal-ct)[dataset]. *The Cancer Imaging Archive*, 2022.
- 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.
- 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):1–9, 2018.
- 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.
- Aaron Van Den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017.
- Bram van Ginneken. Lola11, April 2021. URL <https://doi.org/10.5281/zenodo.4708800>.
- 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.
- Mart van Rijthoven, Witali Aswolinskiy, Leslie Tessier, and Francesco Ciompi. Tiger training dataset (roi-level annotations of wsirois subset), 2022. URL <https://zenodo.org/record/6014422>.
- Maya Varma, Ashwin Kumar, Rogier van der Sluijs, Sophie Ostmeier, Louis Blankemeier, Pierre Chambon, Christian Bluethgen, Jip Prince, Curtis Langlotz, and Akshay Chaudhari. MedVAE: Efficient automated interpretation of medical images with large-scale generalizable autoencoders. *arXiv:2502.14753*, 2025.



- Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, Simon Graham, Quoc Dang Vu, Mieke Zwager, Shan E. Ahmed Raza, Nasir Rajpoot, Xiyi Wu, Huai Chen, Yijie Huang, Lisheng Wang, Hyun Jung, G. Thomas Brown, Yanling Liu, Shuolin Liu, Seyed Alireza Fatemi Jahromi, Ali Asghar Khani, Ehsan Montahaei, Mahdiah Soleymani Baghshah, Hamid Behroozi, Pavel Semkin, Alexandr Rassadin, Prasad Dutande, Romil Lodaya, Ujjwal Baid, Bhakti Baheti, Sanjay Talbar, Amirreza Mahbod, Rupert Ecker, Isabella Ellinger, Zhipeng Luo, Bin Dong, Zhengyu Xu, Yuehan Yao, Shuai Lv, Ming Feng, Kele Xu, Hasib Zunair, Abdessamad Ben Hamza, Steven Smiley, Tang-Kai Yin, Qi-Rui Fang, Shikhar Srivastava, Dwarikanath Mahapatra, Lubomira Trnavska, Hanyun Zhang, Priya Lakshmi Narayanan, Justin Law, Yinyin Yuan, Abhiroop Tejomay, Aditya Mitkari, Dinesh Koka, Vikas Ramachandra, Lata Kini, and Amit Sethi. Monusac2020: A multi-organ nuclei segmentation and classification challenge. *IEEE Transactions on Medical Imaging*, 40(12):3413–3423, 2021.
- Iris Vos, Ynte Ruigrok, and Hugo Kuijf. Results of the crown challenge on automated assessment of circle of willis morphology. In *Medical Imaging with Deep Learning*, 2024.
- Chuanbo Wang, Amirreza Mahbod, Isabella Ellinger, Adrian Galdran, Sandeep Gopalakrishnan, Jeffrey Niezgoda, and Zeyun Yu. Fuseg: The foot ulcer segmentation challenge. *Information*, 15(3):140, 2024a.
- Dequan Wang, Xiaosong Wang, Lilong Wang, Mengzhang Li, Qian Da, Xiaoqiang Liu, Xiangyu Gao, Jun Shen, Junjun He, Tian Shen, et al. A real-world dataset and benchmark for foundation model adaptation in medical image classification. *Scientific Data*, 10(1):574, 2023.
- Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):19549, Nov 2020.
- Wenxuan Wang, Fan Zhang, Yufeng Cui, Haiwen Diao, Zhuoyan Luo, Huchuan Lu, Jing Liu, and Xinlong Wang. End-to-end vision tokenizer tuning. *arXiv preprint arXiv:2505.10562*, 2025.
- 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*, pp. 2097–2106, 2017a.
- 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*, pp. 2097–2106, 2017b.
- Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yuezhe Wang, Zhen Li, Qiyang Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.
- Yuran Wang, Zhijing Wan, Yansheng Qiu, and Zheng Wang. Devil is in details: Locality-aware 3d abdominal ct volume generation for self-supervised organ segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10640–10648, 2024c.
- Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, volume 2, pp. 1398–1402, 2003.
- Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- 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.

- William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian. Breast cancer wisconsin (diagnostic) data set. [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)), 1995. Accessed: 2025-05-22.
- 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*, pp. 87–97, 2024.
- Junde Wu, Huihui Fang, Fei Li, Huazhu Fu, Fengbin Lin, Jiongcheng Li, Yue Huang, Qinji Yu, Sifan Song, Xinxing Xu, et al. Gamma challenge: glaucoma grading from multi-modality images. *Medical Image Analysis*, 90:102938, 2023a.
- Ruiqi Wu, Chenran Zhang, Jianle Zhang, Yi Zhou, Tao Zhou, and Huazhu Fu. MM-Retinal: Knowledge-enhanced foundational pretraining with fundus image-text expertise. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 722–732, 2024.
- Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. VILA-U: a unified foundation model integrating visual understanding and generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=02haSp0453>.
- 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*, 2023b.
- Yiming Xiao, Hassan Rivaz, Matthieu Chabanas, Maryse Fortin, Ines Machado, Yangming Ou, Mattias P Heinrich, Julia A Schnabel, Xia Zhong, Andreas Maier, et al. Evaluation of mri to ultrasound registration methods for brain shift correction: the curious2018 challenge. *IEEE Transactions on Medical Imaging*, 39(3):777–786, 2019.
- Feng Xu, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, and Mulan Jin. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in Oncology*, pp. 4133, 2021.
- Xuanang Xu, Fugen Zhou, Bo Liu, Dongshan Fu, and Xiangzhi Bai. Efficient multiple organ localization in ct image using 3d region proposal network. *IEEE Transactions on Medical Imaging*, 38(8):1885–1898, 2019.
- Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNISTv2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.
- Jinzhong Yang, Greg Sharp, Harini Veeraraghavan, Wouter Van Elmpt, Andre Dekker, Tim Lustberg, and Mark Gooding. Data from lung ct segmentation challenge 2017 (lctsc), 2017. URL <https://www.cancerimagingarchive.net/collection/lctsc/>.
- Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.
- Fengshou Yin, Jiang Liu, Damon Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, Min Thet Htoo, Zhuo Zhang, Yanwu Xu, Xinting Gao, and Ruchir Srivastava. ACHIKO-I retinal fundus image database and its evaluation on cup-to-disc ratio measurement. In *2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA)*, pp. 224–227, 2013.
- yiweichen04. Retina\_dataset, 2016. URL [https://github.com/yiweichen04/retina\\_dataset](https://github.com/yiweichen04/retina_dataset).
- Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A Ross, and Lu Jiang. Language model beats diffusion - tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=gzqrANCF4g>.

- Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, ParasLakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation. <https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation>, 2019. Kaggle.
- Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. AnyGPT: Unified multimodal LLM with discrete sequence modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9637–9662, August 2024.
- 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, 2023a.
- Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023b.
- Chuang Zhu, Wenkai Chen, Ting Peng, Ying Wang, and Mulan Jin. Hard sample aware noise robust learning for histopathology image classification. *IEEE Transactions on Medical Imaging*, 41(4): 881–894, 2021.
- Lei Zhu, Fangyun Wei, Yanye Lu, and Dong Chen. Scaling the codebook size of VQ-GAN to 100,000 with a utilization rate of 99%. In *Advances in Neural Information Processing Systems*, volume 37, pp. 12612–12635, 2024.
- Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2933–2946, 2018.
- Aneeq Zia, Kiran Bhattacharyya, Xi Liu, Ziheng Wang, Satoshi Kondo, Emanuele Colleoni, Beatrice van Amsterdam, Razeen Hussain, Raabid Hussain, Lena Maier-Hein, et al. Surgical visual domain adaptation: Results from the miccai 2020 surgvisdom challenge. *arXiv preprint arXiv:2102.13644*, 2021.
- Aneeq Zia, Max Berniker, Rogerio Garcia Nespole, 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, Hümeysra 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 Abdalbaki 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 Yamlaoui, Piotr Kalinowski, Dominik Michael, Tim Rădscă, 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.
- M. L. Zuley, R. Jarosz, S. Kirk, Y. Lee, R. Colen, K. Garcia, D. Delbeke, M. Pham, P. Nagy, G. Sevinc, M. Goldsmith, S. Khan, J. M. Net, F. R. Lucchesi, and N. D. Aredes. The cancer genome atlas head-neck squamous cell carcinoma collection (tcga-hnsc), 2016. URL <https://doi.org/10.7937/K9/TCIA.2016.LXKQ47MS>. Data set.

# Appendix of Paper: “MedITok: A Unified Tokenizer for Medical Image Synthesis and Interpretation”

## CONTENTS

<b>A Training Dataset</b>	<b>37</b>
A.1 Data Collection . . . . .	37
A.2 Data Preprocessing . . . . .	37
A.3 Data Statistics . . . . .	38
<b>B Benchmarking Datasets</b>	<b>40</b>
B.1 Image Reconstruction . . . . .	40
B.2 Image Classification . . . . .	40
B.3 Image Synthesis . . . . .	40
B.4 Visual Question Answering . . . . .	40
<b>C Experimental Setups</b>	<b>41</b>
C.1 Implementation Details . . . . .	41
C.2 Downstream Tasks . . . . .	41
<b>D Additional Results</b>	<b>42</b>
D.1 Additional Ablation Studies . . . . .	42
D.2 Adaptation to 3D Medical Volumes . . . . .	44
D.3 Visual Turing Test . . . . .	44
D.4 Difference between Stages . . . . .	45
D.5 Failure Cases . . . . .	46
D.6 Efficiency Comparison . . . . .	46
D.7 Additional Visualization . . . . .	48
<b>E Discussion</b>	<b>48</b>
E.1 Choice of Discrete Codebooks . . . . .	49
E.2 Comparison with Related Works . . . . .	49
E.3 Limitation and Future Directions . . . . .	50
E.4 Broader Impact . . . . .	50



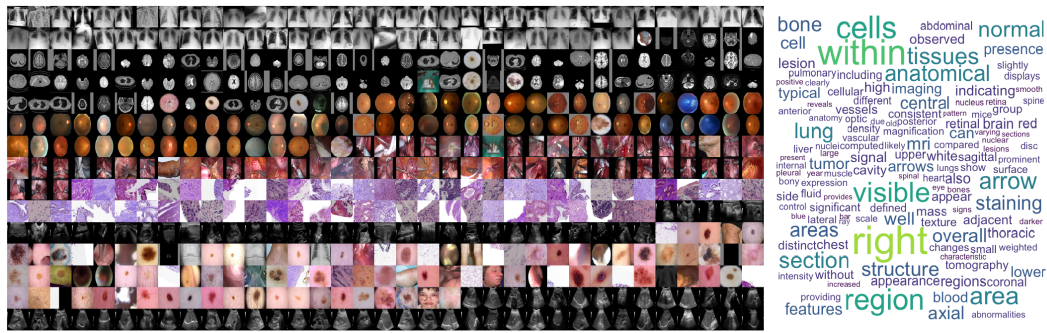


Figure S1: Overview of the training data for MedITok. Left: exemplar images used in the first training stage. Right: word cloud generated from the captions used in the second training stage.

[Return to: Introduction](#) | [Experiments](#) | [Appendix Contents](#)

## A TRAINING DATASET

In this section, we provide a comprehensive overview of the training dataset used in this work, including the collection (Appendix A.1), preprocessing (Appendix A.2), and statistics (Appendix A.3) of image-only datasets and image-text paired datasets. The construction of this training dataset is pivotal to the success of our proposed MedITok, as it ensures a diverse and high-quality representation of medical images and text descriptions across multiple modalities, anatomical regions, and clinical contexts.

### A.1 DATA COLLECTION

We begin by identifying and collecting medical imaging datasets from over 300 publicly available sources, ensuring broad coverage of imaging modalities and clinical scenarios. Our selection criteria include: (1) **Appropriate Licensing**: We only select datasets with clear licensing, ensuring compliance with data usage policies; (2) **Clinical Relevance**: Only datasets that provide diagnostic-quality images or clinically annotated images were included; and (3) **Diversity in Imaging Modalities and Anatomies**: We prioritize datasets that collectively cover a wide range of anatomical regions and pathologies.

### A.2 DATA PREPROCESSING

#### A.2.1 EXTRACTING 2D IMAGES FROM 3D VOLUMES

A significant portion of our dataset comprises volumetric medical images (CT and MRI). To fully utilize these data to train our 2D visual tokenizer, we carefully convert them into 2D image slices using a modality-specific preprocessing strategy.

**CT images extracted from volumes.** Each 3D CT volume is first converted to Hounsfield Units (HU) using the rescaling slope and intercept recorded in the metadata, and is then clipped to the range of  $[-1000, 2000]$ . To obtain 2D slices from the 3D volume, we extract slices along three orthogonal planes (axial, coronal, and sagittal), and select every fifth slice along each plane. We then perform an initial quality filtering by retaining CT slices that met all the following criteria: (1) a background ratio (the proportion of pixels with HU values  $\leq -1000$ )  $\leq 0.6$ , (2) a valid body ratio (the proportion of pixels with HU values  $\geq -300$ )  $\geq 0.1$ , and (3) a pixel intensity standard deviation  $< 100$ . These criteria ensure the removal of largely empty slices with minimal anatomical content.

Note that, we save the CT images extracted from 3D volumes in their original HU values without scaling them to the  $[0, 255]$  range. By doing so, we can apply various CT window settings on the CT images during model training as a form of data augmentation, as detailed in Appendix C.1.

**MRI images extracted from volumes.** We process each MRI volume by clipping voxel values to the  $[0.5^{\text{th}}, 99.5^{\text{th}}]$  percentile range, followed by min-max normalization to  $[-1, 1]$ . The 2D slices are extracted using the same way as CT preprocessing. The initial quality filtering for MRI excludes those slices with mean pixel values  $\leq -0.9$  or standard deviation  $\leq 0.2$ .

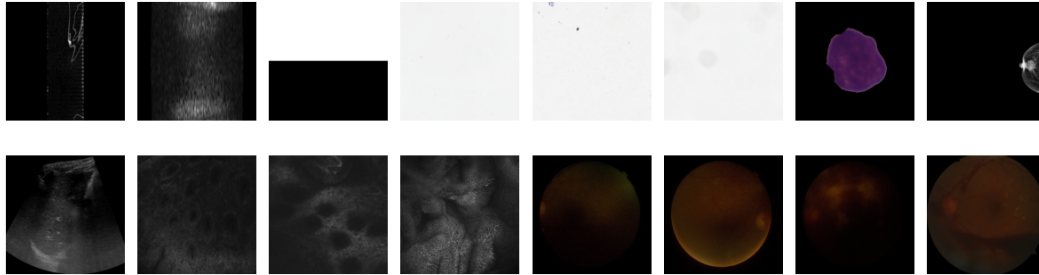


Figure S2: Low-quality images filtered by our quality control pipeline.

### A.2.2 QUALITY CONTROL

Once we obtain all the 2D images, we implement the following process to ensure that only high-quality, clinically relevant images are included in the training dataset:

- **Dynamic Range Check:** Images with pixel intensity ranges below 50 were filtered out to ensure adequate contrast.
- **Resolution Filtering:** Images with a minimum dimension below 128 pixels were excluded to maintain structural integrity.
- **Information Content Validation:** Images with low standard deviation (below 10) in pixel values were discarded.
- **Palette Limitation Removal:** Images with three or fewer unique pixel values were removed.
- **Relevance Verification:** Non-clinical images, such as tables, plots, or irrelevant illustrations, were manually screened and excluded.

For instance, the “Relevance Verification” is mainly applied on the BIOMEDICA (Lozano et al., 2025) dataset, which originally contains approximately 24,050,423 image-text pairs extracted from biomedical publications. Each image-text pair is tagged with primary and secondary labels. We retain only those pairs with a primary label of “Clinical Imaging” and a secondary label matching one of the following: “computerized tomography”, “clinical imaging”, “light microscopy”, “immunohistochemistry”, “endoscopy”, “eye”, “X-ray radiography”, “ultrasound”, “magnetic resonance”, “brain”, “skin lesion”, and “mammography”. Image-text pairs tagged with irrelevant secondary labels (e.g., “scientific illustration”, “ambiguous”, “plot”, “diagram”, *etc.*) are all excluded. Such filtering significantly reduces the BIOMEDICA dataset from 24,050,423 to 1,216,529 image-text pairs for use in our experiments.

Following the automated checks described above, we perform a manual quality assessment by randomly sampling 10 images from each dataset for manual visual inspection. If any low-quality outliers are identified, we further examine other images from the corresponding dataset to evaluate overall quality. Finally, we try our best to remove the images that share the same sources with data in our benchmarking datasets in Appendix B.

Fig. S2 displays some low-quality images detected by the dynamic range check, information content validation, and palette limitation removal. For another example, Fig. S3 shows images that are not tagged as “clinical imaging” in the original BIOMEDICA (Lozano et al., 2025) dataset.

### A.3 DATA STATISTICS

After the collection and the preprocessing, we present detailed sources and image counts of our “image-only” dataset, which is used in the first training stage of MedITok, in Tables S5–S8. The details of the “image-caption” dataset, used in the second training stage, are presented in Table S9.

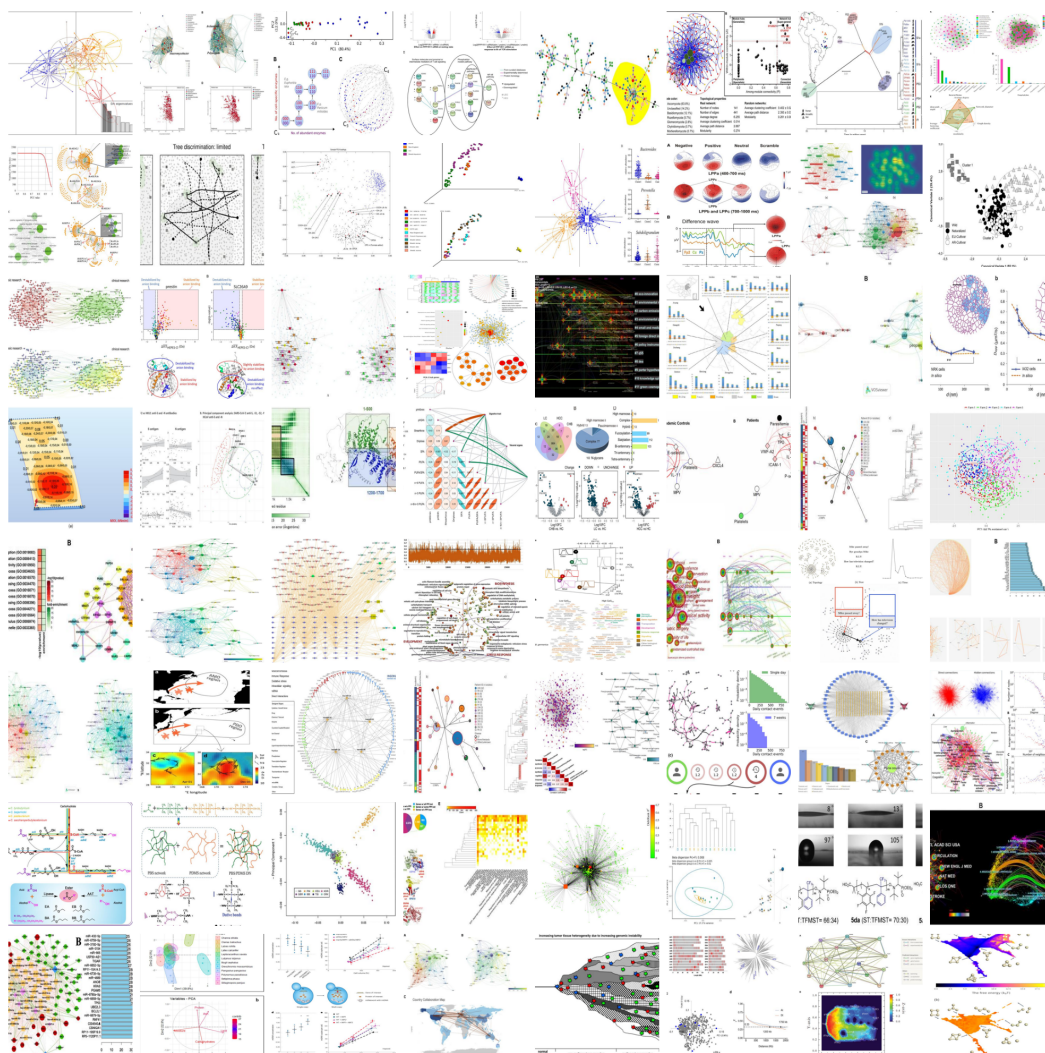


Figure S3: Irrelevant images filtered out by our quality control pipeline.

Return to: [Introduction](#) | [Experiments](#) | [Appendix Contents](#)

## B BENCHMARKING DATASETS

This section outlines the datasets used for evaluating the performance of MedITok across four core tasks: medical image reconstruction (Appendix B.1), classification (Appendix B.2), modality-conditioned image synthesis (Appendix B.3), and visual question answering (Appendix B.4). We tried our best to avoid any overlap or data leakage between the training data of MedITok and these benchmark datasets.

### B.1 IMAGE RECONSTRUCTION

To assess the reconstruction capabilities of MedITok, we curated a high-quality evaluation set of 35,736 images spanning 8 imaging modalities. These images are collected from 23 publicly available datasets, as detailed in Table S10. Importantly, all images used for evaluation are strictly excluded from the training corpus to prevent any overlap. This evaluation set reflects a diverse distribution of anatomical structures, imaging protocols, and clinical contexts, enabling robust testing of image fidelity and structural preservation.

### B.2 IMAGE CLASSIFICATION

We adopt five subsets from the latest version<sup>2</sup> of MedMNIST (Yang et al., 2023) benchmark to evaluate the semantic encoding quality of the visual tokens produced by different tokenizers. These include:

- BreastMNIST (Al-Dhabyani et al., 2020) (ultrasound): binary classification of benign vs. malignant tumors.
- DermaMNIST (Tschandl et al., 2018; Codella et al., 2019) (dermoscopy): 7-way classification of skin lesions.
- PathMNIST (Kather et al., 2019) (pathology): 9-class colorectal cancer tissue types.
- PneumoniaMNIST (Kermany et al., 2018) (X-ray): pneumonia detection in chest radiographs.
- RetinaMNIST (Liu et al., 2022) (fundus): diabetic retinopathy grading.

The original images in each benchmark are of a resolution of  $224 \times 224$ , and are resized to  $256 \times 256$  resolution for consistency with the tokenizer input. These tasks collectively test the extent to which the visual tokenizer encodes discriminative, clinically meaningful semantics. Detailed training and test split can be found in Table S11.

### B.3 IMAGE SYNTHESIS

To evaluate the generative capability of downstream autoregressive models built on top of MedITok, we conduct experiments on modality-conditioned image synthesis. Specifically, we use six subsets from the latest MedMNIST collection (Yang et al., 2023), including BloodMNIST (Acevedo et al., 2020) for microscopy, BreastMNIST (Al-Dhabyani et al., 2020) for ultrasound, ChestMNIST (Wang et al., 2017b) for chest x-ray, DermaMNIST (Tschandl et al., 2018; Codella et al., 2019) for dermoscopy, PathMNIST (Kather et al., 2019) for pathology images, and RetinaMNIST (Liu et al., 2022) for fundus photography. We gather the training partition of these subsets with their imaging modality labels to construct the training data for the downstream medical image synthesis models, which are trained to generate images conditioned on modality labels.

### B.4 VISUAL QUESTION ANSWERING

To test the utility of different visual tokenizers for medical image interpretation in multimodal settings, we benchmark on two widely adopted datasets for visual question answering (VQA) task:

<sup>2</sup><https://doi.org/10.5281/zenodo.10519652>



(1) VQA-RAD (Lau et al., 2018): A radiology-specific VQA dataset with natural language questions and answers grounded in diagnostic images. We use its test set containing 451 question-answer pairs for evaluation. (2) SLAKE (Liu et al., 2021a): A multi-modal, bilingual medical VQA benchmark with more diverse imaging modalities and question types. The validation set (SLAKE-val) with 2,094 questions and test set (SLAKE-test) with 2,099 questions are adopted in our experiments.

To train vision-language model for medical image interpretation (*i.e.*, LLaVA-Med (Li et al., 2023) variants with different visual tokenizers as the image encoder), we use the PubMedVision (Chen et al., 2024c) dataset, which consists of high-quality image-question-answer triplets derived from medical publications. All VQA benchmarks are held out from the training set to ensure fair and unbiased evaluation.

Return to: [Introduction](#) | [Experiments](#) | [Appendix Contents](#)

## C EXPERIMENTAL SETUPS

In this section, we first describe the detailed implementation and training setup of MedITok (Appendix C.1) and its downstream applications (Appendix C.2) on four core tasks: image reconstruction, image classification, image synthesis, and visual question answering.

### C.1 IMPLEMENTATION DETAILS

**Architecture.** MedITok consists of a ViTamin-based (Chen et al., 2024b) image encoder and decoder, with a multi-codebook vector quantizer (Ma et al., 2025b) in the bottleneck. The encoder produces a 2D grid of latent representations, which are discretized using 8 parallel codebooks, each with 4,096 eight-dimensional vectors, resulting in a total vocabulary size of 32,768. The decoder reconstructs the image from quantized latent vectors.

**Training of MedITok.** Both training stages (*i.e.*, visual representation alignment, and textual semantic alignment) share the same reconstruction loss defined as follows:

$$\mathcal{L}_{\text{recon}}(\hat{\mathbf{x}}, \mathbf{x}, \mathbf{z}_q, \mathbf{z}) = \mathcal{L}_{\text{image}}(\hat{\mathbf{x}}, \mathbf{x}) + \lambda_{\text{comm}} \mathcal{L}_{\text{comm}}(\mathbf{z}_q, \mathbf{z}), \quad (\text{S1})$$

$$\mathcal{L}_{\text{image}}(\hat{\mathbf{x}}, \mathbf{x}) = \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}(\hat{\mathbf{x}}, \mathbf{x}) + \lambda_{\text{perc}} \mathcal{L}_{\text{perc}}(\hat{\mathbf{x}}, \mathbf{x}), \quad (\text{S2})$$

$$\mathcal{L}_{\text{comm}}(\mathbf{z}_q, \mathbf{z}) = \|\mathbf{z}_q - \text{sg}[\mathbf{z}]\|_2^2 + \beta \|\text{sg}[\mathbf{z}_q] - \mathbf{z}\|_2^2, \quad (\text{S3})$$

where  $\mathcal{L}_{\text{adv}}$  denotes the adversarial loss (Esser et al., 2021),  $\mathcal{L}_{\text{perc}}$  the perceptual loss (Johnson et al., 2016), and  $\mathcal{L}_{\text{comm}}$  the commitment loss (Van Den Oord et al., 2017). “sg[.]” denotes the stop-gradient operation. We follow the default setting of VQGAN (Esser et al., 2021) to set  $\lambda_{\text{adv}}$  as an adaptive weight and fix  $\beta = 0.25$ ,  $\lambda_{\text{perc}} = 1$ , and  $\lambda_{\text{comm}} = 1$ . The discriminator involved in computing  $\mathcal{L}_{\text{adv}}$  adopts the DINOv2 (Oquab et al., 2023) architecture. We use the AdamW (Loshchilov & Hutter, 2019) optimizers for both MedITok and the discriminator, with betas of (0.9, 0.95) and a weight decay of 0.02 for MedITok, and (0.5, 0.9) and 0.2 for the discriminator. The learning rate is initialized at  $5 \times 10^{-4}$  and decayed to  $5 \times 10^{-5}$  via cosine annealing; for the discriminator, it starts at  $2 \times 10^{-5}$  and decays to  $2 \times 10^{-6}$ . The two-stage full-data training took approximately 4 days on 8 NVIDIA H100 GPUs.

We employ random resized cropping, random image flipping, random image rotation for data augmentation in the first training stage. For CT image input in HU values, we further introduce **CT windowing augmentation**, which randomly applies the following windows on the HU values: full window  $[-1000, 2000]$  HU, probability  $p = 0.2$ , common window  $[-1000, 1000]$  HU,  $p = 0.3$ , soft tissue window  $[-150, 250]$  HU,  $p = 0.3$ , lung window  $[-1400, 200]$  HU,  $p = 0.15$ , and bone window  $[-500, 1300]$  HU,  $p = 0.05$ .

### C.2 DOWNSTREAM TASKS

**Medical image classification.** For classification tasks, we evaluate the discriminative power of the learned visual tokens through a linear probing protocol (Alain & Bengio, 2016). Specifically, for a pretrained visual tokenizer (*e.g.*, MedITok), we only use its image encoder and quantizer, keep

them frozen, and append a single linear layer on top of the quantizer. Given an image, the image encoder produces the continuous feature maps, which are then discretized to a grid of visual tokens and are flattened to feed the linear layer for image classification. The linear layer is trained using the Adam (Kingma, 2014) optimizer with a learning rate of  $10^{-4}$  and a batch size of 128. Since the tokens produced by different tokenizers lead to different convergence speed for the linear layer, we train each linear layer until convergence and report the peak performance for a fair comparison.

**Medical image synthesis.** For image synthesis, we integrate the visual tokenizer with LlamaGen-B (Sun et al., 2024), an autoregressive model designed for image generation, with 12 transformer layers, 12 heads, and 768 token dimension. We first tokenize each training image, producing a discrete token sequence. Then, LlamaGen is trained to autoregressively predict the token sequence conditioned on a modality label token. LlamaGen models are optimized using AdamW (Loshchilov & Hutter, 2019) with betas of (0.9, 0.95), a weight decay of 0.05, and a learning rate of  $10^{-4}$ . The models are trained for 200 epochs with a batch size of 128. We do not employ advanced strategy for sampling (e.g., classifier-free guidance) and synthesize images with a temperature parameter of 1.

**Medical visual question answering.** For VQA, we adapt LLaVA-Med (Li et al., 2023) by replacing its image encoder with different visual tokenizers, followed by a projector to produce visual embeddings compatible with the pretrained language backbone<sup>3</sup>. We follow the staged training procedure of original LLaVA-Med, which includes a pretraining stage for the projector (with all other components frozen) and a fine-tuning stage for the language model using LoRA (Hu et al., 2022). The pretraining is conducted on 500k image-caption pairs in PubmedVision (Chen et al., 2024c) dataset for one epoch with batch size 32, while the fine-tuning takes two epochs on the 100k visual question-answer pairs.

Return to: [Introduction](#) | [Experiments](#) | [Appendix Contents](#)

## D ADDITIONAL RESULTS

This section compiles extended evidence to complement the main results, including additional ablation studies (Appendix D.1), adaptation to 3D medical volumes (Appendix D.2), additional visual Turing test for medical image synthesis (Appendix D.3), analyses of the differences between codebooks across training stages (Appendix D.4) and representative failure cases (Appendix D.5), comparison of data scale and inference efficiency (Appendix D.6), and additional visualizations for reconstruction, synthesis, and VQA (Appendix D.7), including qualitative generative and VQA examples that illustrate behavior beyond aggregate metrics.

### D.1 ADDITIONAL ABLATION STUDIES

We present additional ablation studies in Table S1 to further investigate the effectiveness of our data quality control and the proposed training framework.

**Separating Rather Than Combining Two Stages.** In contrast to previous works, we propose incorporating an extra training stage (e.g., visual representation alignment) in the training of unified visual tokenizer. A natural idea question comes: can we combine this stage and the subsequent textual semantic alignment stage together in one stage? That is, in each iteration, we optimize the following loss function:

$$\begin{aligned}\mathcal{L} = & \mathcal{L}_{\text{recon}}(\hat{x}, x, z_q, z) \\ & + \lambda_{\text{vision}} \mathcal{L}_{\text{vision}}(z_q, f_{\text{vision}}(\mathcal{E}_{\text{vision}}(x))) \\ & + \lambda_{\text{text}} \mathcal{L}_{\text{text}}(z_q, f_{\text{text}}(\mathcal{E}_{\text{text}}(t))),\end{aligned}\tag{S4}$$

In Rows (i) and (ii) of Table S1, we empirically compare combined single-stage and our two-stage training under the same setting. The combined-stage training only slightly improves semantic metrics but significantly degrades reconstruction quality. This may be attributed to the dominance of semantic alignment objectives, which in turn escalates the inherent conflicts between reconstruction (low-level) and semantic (high-level) alignment objectives. In contrast, we use the visual representation learning

<sup>3</sup><https://huggingface.co/microsoft/llava-med-v1.5-mistral-7b>

Table S1: More ablation studies of MedITok. “#Img”: number of images used in the first training stage, “#Img-txt”: number of image-text pairs used in the second training stage. “BiomedCLIP-T (combined)”: textual semantic alignment is combined with the visual representation alignment as one single stage. “BiomedCLIP-T<sup>†</sup>”: the BiomedCLIP (Zhang et al., 2023b) text encoder is activated during training.

Idx.	Vision Target Repr.	Text Target Repr.	$\lambda_{\text{vision}}$	#Img	#Img-txt	PSNR	SSIM	mAP	AUC
(i)	BiomedCLIP-V	BiomedCLIP-T (combined)	0.1	800k	1M	29.20	83.22	81.10	91.97
(ii)	BiomedCLIP-V	BiomedCLIP-T	0.1	800k	1M	30.03	84.32	80.09	92.64
(iii)	–	BiomedCLIP-T	0	800k	24M (all BIOMEDICA)	32.23	89.36	57.97	76.98
(iv)	–	BiomedCLIP-T	0	800k	1M (filtered BIOMEDICA)	32.55	89.49	63.29	81.68
(v)	BiomedCLIP-V	BiomedCLIP-T	0.1	800k	1M	30.03	84.32	80.09	92.64
(vi)	BiomedCLIP-V	BiomedCLIP-T	1	800k	1M	29.99	83.02	82.00	91.81
(vii)	–	BiomedCLIP-T	0	–	2.4M	29.06	79.61	80.29	91.25
(viii)	–	BiomedCLIP-T	0	–	2.4M (+800k)*	30.05	82.12	80.06	91.18
(ix)	BiomedCLIP-V	BiomedCLIP-T	0.1	800k	2.4M	29.74	84.14	80.28	92.72
(x)	BiomedCLIP-V	BiomedCLIP-T	0.1	2M	2.4M	30.20	85.50	82.23	93.61
(xi)	BiomedCLIP-V	BiomedCLIP-T <sup>†</sup>	0.1	33.4M	2.4M	34.03	91.05	51.41	69.84
(xii)	BiomedCLIP-V	BiomedCLIP-T	0.1	33.4M	2.4M	31.74	88.25	82.27	94.07
(xiii)	BiomedCLIP-V	–	Cos. sim	800k	–	30.18	84.01	66.19	85.77
(xiv)	BiomedCLIP-V	–	Contrast	800k	–	30.00	83.85	78.35	92.23

\*: we convert 800k pure images to pseudo image-text pairs by tagging each image with a short caption “This is a \${modality} image.”

as a cold-start to transit from a reconstruction-based tokenizer to a unified tokenizer more smoothly, improving joint optimization of these competing objectives.

We also note that separating two stages provides more flexibility, particularly when training with significantly imbalanced data collections in the medical domain, where unlabeled images are far more abundant than image-text pairs (14x in our final training set). A staged design allows us to exploit such imbalanced data effectively and provides engineering flexibility for making modifications to the pretrained encoders (*e.g.*, adding trainable parameters), while avoiding potential gradient issues caused by heterogeneous batches.

**Data Quality Control.** Rows (iii) and (iv) of Table S1 presents the result from our pilot study to evaluate the effectiveness of our data quality control pipeline. We pretrain MedITok with pure reconstruction objective in the first training stage, and continue the second training stage on the BIOMEDICA (Lozano et al., 2025) dataset.

Specifically, in Row (iii), we adopt all 24M image-text pairs in this dataset, while in Row (iv), we use a much smaller subset with approximately 1M pairs, as described in Appendix A.2.2. Surprisingly, despite the significant reduction in the training dataset size, the tokenizer in Row (iv) exhibits much stronger medical image reconstruction and classification capabilities, compared to the one in Row (iii). This highlights the importance of data quality control in training a powerful visual tokenizer<sup>4</sup>.

$\lambda_{\text{vision}}$  **Balancing Reconstruction and Contrastive Learning.** In Rows (v) and (vi), we explore the effect of different magnitude for the visual representation alignment in the first training stage by varying  $\lambda_{\text{vision}}$  in Eq. 2. By setting a light semantic constraint ( $\lambda_{\text{vision}} = 0.1$ ), we observe an improvement across three metrics (PSNR, SSIM, and AUC) while maintaining competitive mAP, and we fix this factor in other experiments.

**Cold-Starting with Visual Representation Alignment.** In Rows (vii) and (viii), we bypass the visual representation alignment stage and train MedITok solely using the textual semantic alignment objective. While this configuration yields reasonable semantic performance, it exhibits a significant degradation in SSIM, compared with other configurations like Row (ix) of Table S1, showing the necessity of the visual pretraining stage for cold-starting MedITok by learning structural coherent representations with a light semantic constraint.

<sup>4</sup>We would like to note that this filtering was tailored to downstream tasks that mainly involve clinical images, and that other image types (*e.g.*, tables, plots, and non-clinical images) in BIOMEDICA remain highly valuable for applications that require table understanding or scientific figure interpretation.

**Freezing the Pretrained Text Encoder.** In Row (xi), we investigate the impact of unfreezing the BiomedCLIP text encoder during the second stage. Although this introduces learnable capacity into the text encoder, it disrupts the stability and alignment of the token space, leading to a trade-off: improved reconstruction metrics but severely degraded downstream classification, compared to the results in Row (xii). This suggests that freezing the pretrained textual backbone acts as an anchor, preserving the semantic information necessary for clinical interpretation.

**Visual Representation Alignment Objective.** We explore two alignment objectives for training MedITok: contrastive learning and cosine similarity (inspired by Yao et al. (2025)). Comparing Rows (xiii) and (xiv), we observe that using cosine-similarity loss yields only marginal gains in PSNR but substantially degrades downstream classification, whereas the contrastive objective produces a more discriminative token space, improving both fine-grained classification and maintaining reconstruction quality.

## D.2 ADAPTATION TO 3D MEDICAL VOLUMES

Three-dimensional data are vital in the medical domain. Our initial milestone targeted a 2D image tokenizer, considering that (1) 2D images cover more medical imaging domains (*e.g.* fundus photography, histopathology, *etc.*), (2) 2D models provide more flexibility, and (3) computational costs.

However, we note that MedITok can also be applied in 3D medical data. We compare MedITok, UniTok, and MedVAE on two 3D datasets: SLIVER07 (Heimann et al., 2009) for volume reconstruction and OrganMNIST3D (Bilic et al., 2023; Xu et al., 2019) for multi-class volume classification of 11 body organs. To adapt these 2D tokenizers to 3D volumes, we employed a slice-based strategy: processing individual slices independently and then aggregating either reconstructed slices (for reconstruction) or per-slice features (for classification). The results are summarized in Table S2.

Table S2: Additional evaluation on 3D datasets.

Models	rFID	PSNR	SSIM	mAP	AUC
MedVAE	20.38	<b>34.21</b>	<b>89.98</b>	76.04	94.77
UniTok	6.89	31.08	86.16	<u>83.25</u>	<u>96.15</u>
MedITok	<b>4.94</b>	<u>33.56</u>	<u>89.54</u>	<b>84.00</b>	<b>97.71</b>

Despite not being trained explicitly on 3D radiology data, MedITok still achieves reconstruction quality comparable to MedVAE which is a *radiology-specialized* visual tokenizer, with notably lower rFID for better visual fidelity and competitive PSNR/SSIM indicating reconstruction accuracy. UniTok encodes visual semantics, yet failing to preserve critical structural details with a significant drop in PSNR and SSIM. More importantly, MedITok significantly outperforms MedVAE on 3D volume classification tasks, proving superior transferable representations in 3D settings. Visualization of 3D reconstruction results are shown in Fig. S4.

## D.3 VISUAL TURING TEST

We conducted a Visual Turing Test on the downstream medical image synthesis task, as a proxy evaluation of the quality of latent space encoded by different tokenizers. Specifically, we compare two autoregressive medical image synthesis models as in Sec. 4.4: (1) LlamaGen-MedITok, using MedITok as its visual tokenizer; and (2) LlamaGen-UniTok, using UniTok instead, a state-of-the-art unified visual tokenizer.

We randomly mixed 75 chest X-rays: 25 real, 25 synthesized by LlamaGen-MedITok, and 25 by LlamaGen-UniTok, and asked a board-certified radiologist with over 10 years’ experience to score the “realness” of each image on a continuous 0–1 scale. From these scores, we computed (i) AUC for classifying real versus synthetic images and (ii) “fooling rate” or “over-confidence”, the proportion of synthetic images scored higher than 0.5. As shown in Table S3, the radiologist had more difficulty distinguishing MedITok-synthesized images from real ones, indicating that MedITok enables a more clinically plausible latent space.



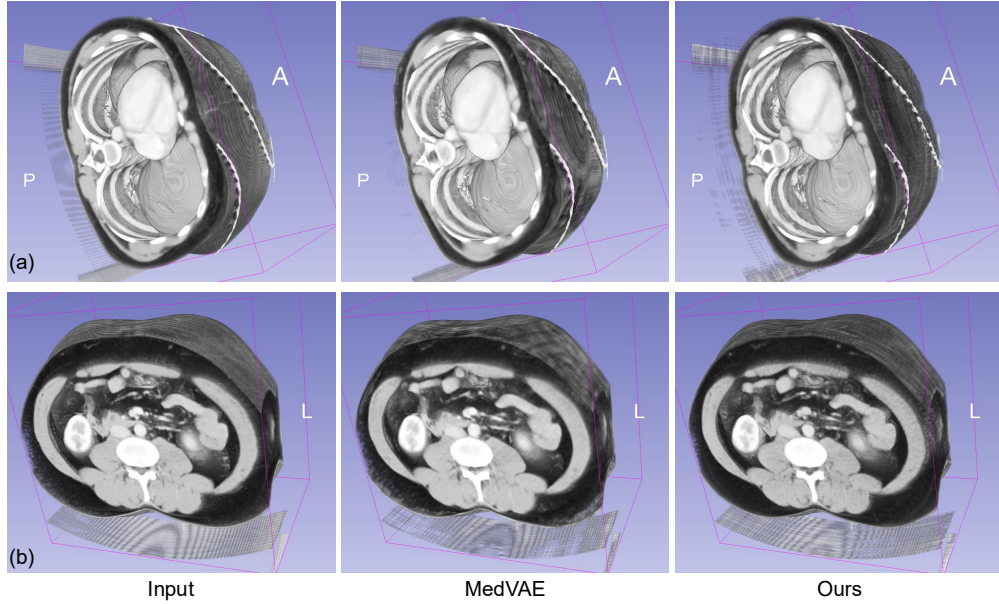


Figure S4: Visualization of 3D reconstruction results.

Table S3: Visual Turing test on downstream medical image synthesis. We report area under the ROC curve (AUC) for real vs. synthetic discrimination and the fooling rate.

Model	AUC	Fooling rate
LlamaGen-UniTok	0.602 (95% CI 0.430–0.772)	56.0% (CI 37.1–73.3%)
LlamaGen-MedITok	<b>0.462</b> (95% CI 0.307–0.622)	<b>72.0%</b> (CI 52.4–85.7%)

#### D.4 DIFFERENCE BETWEEN STAGES

We compare the two stages through both performance behavior and the geometry of their learned codebooks. In Tables 3 and 4, models built upon the Stage-2 MedITok (4th row) significantly outperform those using the Stage-1 tokenizer (3rd row) in both image synthesis and interpretation, confirming that Stage 2 enhances semantic capacity without sacrificing reconstruction quality.

Empirically, Fig. S5 shows the test performance curve. In Stage 1, rFID steadily decreases while mAP remains flat or drifts slightly downward, consistent with a phase that emphasizes reconstructive accuracy over discriminative semantics. When training continues into Stage 2, mAP rises sharply, showing a strong boost in classification performance as semantic constraints are reinforced. rFID exhibits a transient increase at the first epoch in Stage 2 but then returns to a level close to the endpoint of Stage 1, indicating that reconstruction quality is largely preserved. Overall, these dynamics support the design of the two-stage schedule: Stage 1 secures a high-fidelity codebook with light semantic constraint, and Stage 2 further enhances clinical semantics in the latent vectors while retaining structural information encoding.

To understand why, we visualize the codebook vectors with *t*-SNE. As shown in Fig. S6, after Stage 2 (strong semantic alignment), the vectors spread more uniformly, pushing features to be well-distributed on the hypersphere, whereas Stage 1 (light semantic constraint) produces visibly clustered pockets.

The clustering in Stage 1 is also consistent with known VQ-VAE behavior: without additional pressures, codebooks tend to exhibit codebook collapse (Roy et al., 2018), yielding concentrated regions in latent space. The move toward a more uniform, semantically aligned latent in Stage 2 therefore explains both the stronger interpretive/synthesis performance. Notably, recent work (Yao et al., 2025) in latent diffusion reaches a congruent conclusion: aligning VAE latents to semantic-rich

features promotes generative quality by regularizing the latent geometry, with only limited impact on reconstruction.

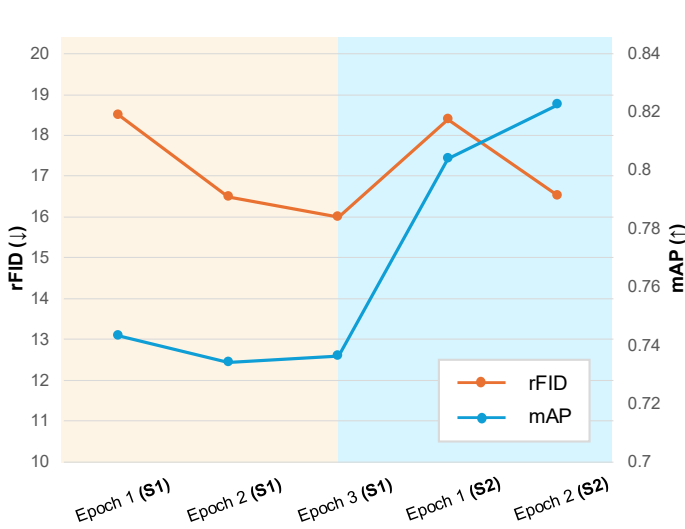


Figure S5: Training dynamics of MedITok, where rFID on the reconstruction test set and mAP on the classification test set are reported for checkpoints from three Stage-1 (S1) epochs followed by two Stage-2 (S2) epochs.

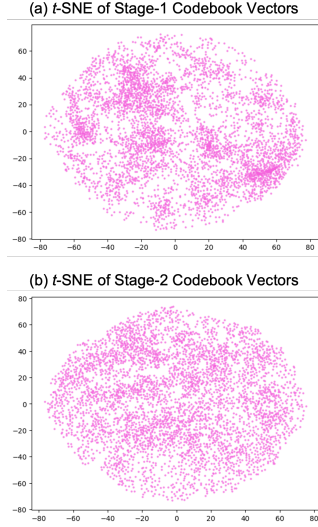


Figure S6: *t*-SNE visualization of codebook vectors in two training stages.

## D.5 FAILURE CASES

Despite the inspiring performance, MedITok may produce inferior reconstruction for histopathology images due to their rich fine-grained textures and structural complexity. As shown in the “Patho.” column of Table 1, all tokenizers struggle with this modality, though MedITok still outperforms existing baselines. This represents a common challenge in histopathology tokenization that warrants future investigation. Qualitative examples for these failure cases are shown in Fig. S7.

## D.6 EFFICIENCY COMPARISON

In Table S4, we provide details on the inference GPU memory consumption (GB), and frame-per-second (FPS) throughput across different settings (*e.g.*, B8: batch size 8, R256: resolution 256). MedITok achieves comparable memory consumption and throughput to existing tokenizers while delivering state-of-the-art reconstruction quality and latent representation (Tables 1 and 2), showing both efficiency and effectiveness.

Table S4: Comparison of different models in terms of inference memory usage, and frames per second (FPS).

Model	Memory (B16, R256)	Memory (B8, R512)	FPS (B16, R256)	FPS (B8, R512)
VQGAN	3.29	6.31	136.24	17.76
PUMIT	0.36	0.56	4440.09	1691.37
VAR-VQ	4.21	7.97	171.26	40.95
Emu3-VQ	41.12	OOM	12.68	OOM
VAR-VQ	4.21	7.98	171.26	40.95
TokenFlow	7.91	Not Supported	44.15	Not Supported
MedVAE	4.61	8.89	101.56	24.34
MedITok	4.69	6.75	92.81	20.63

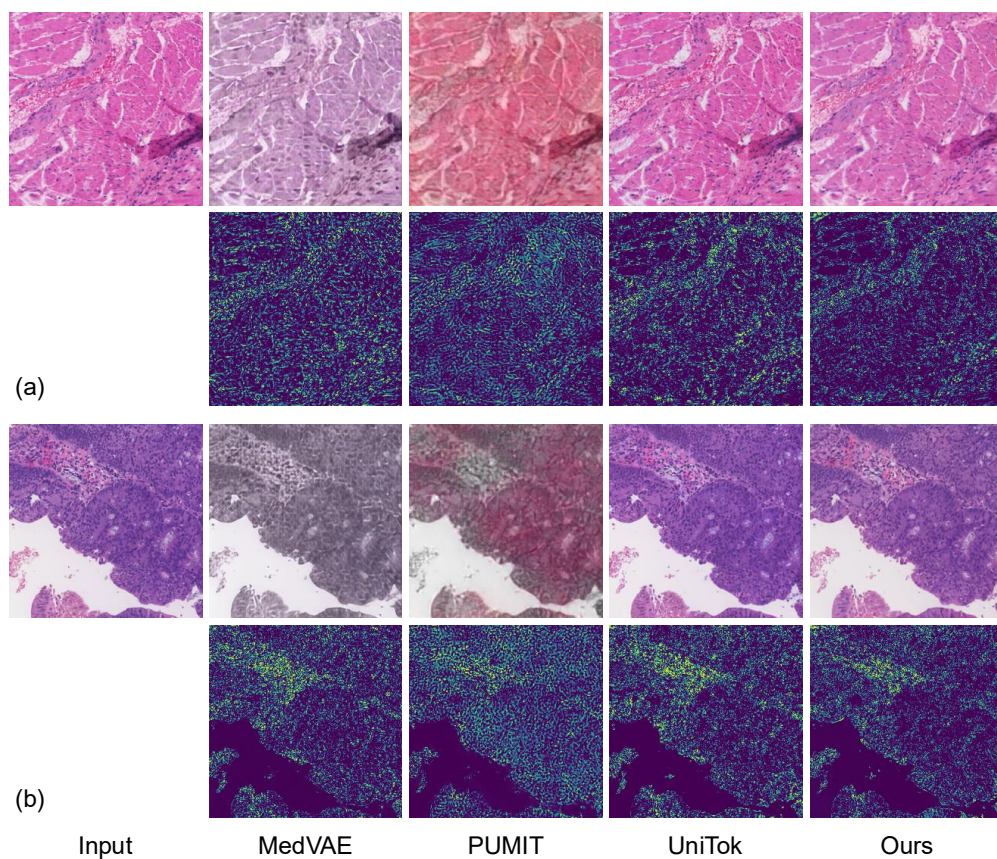


Figure S7: Two failure cases for image reconstruction. Due to the extremely rich details in histopathology images, existing visual tokenizers may still produce lower-fidelity reconstructions.



### D.7 ADDITIONAL VISUALIZATION

Fig. S8 shows more examples for qualitative evaluation of medical image reconstruction, where MedITok achieves the best visual quality with lowest errors. Fig. S9 compares the modality-conditioned synthesized images produced by different LlamaGen models. Notably, the LlamaGen model that adopts our MedITok as the visual tokenizer yields diverse and realistic medical images. Figs. S10–S12 presents the visual question answering results of LLaVA models that incorporate different visual tokenizers as their respective image encoder.

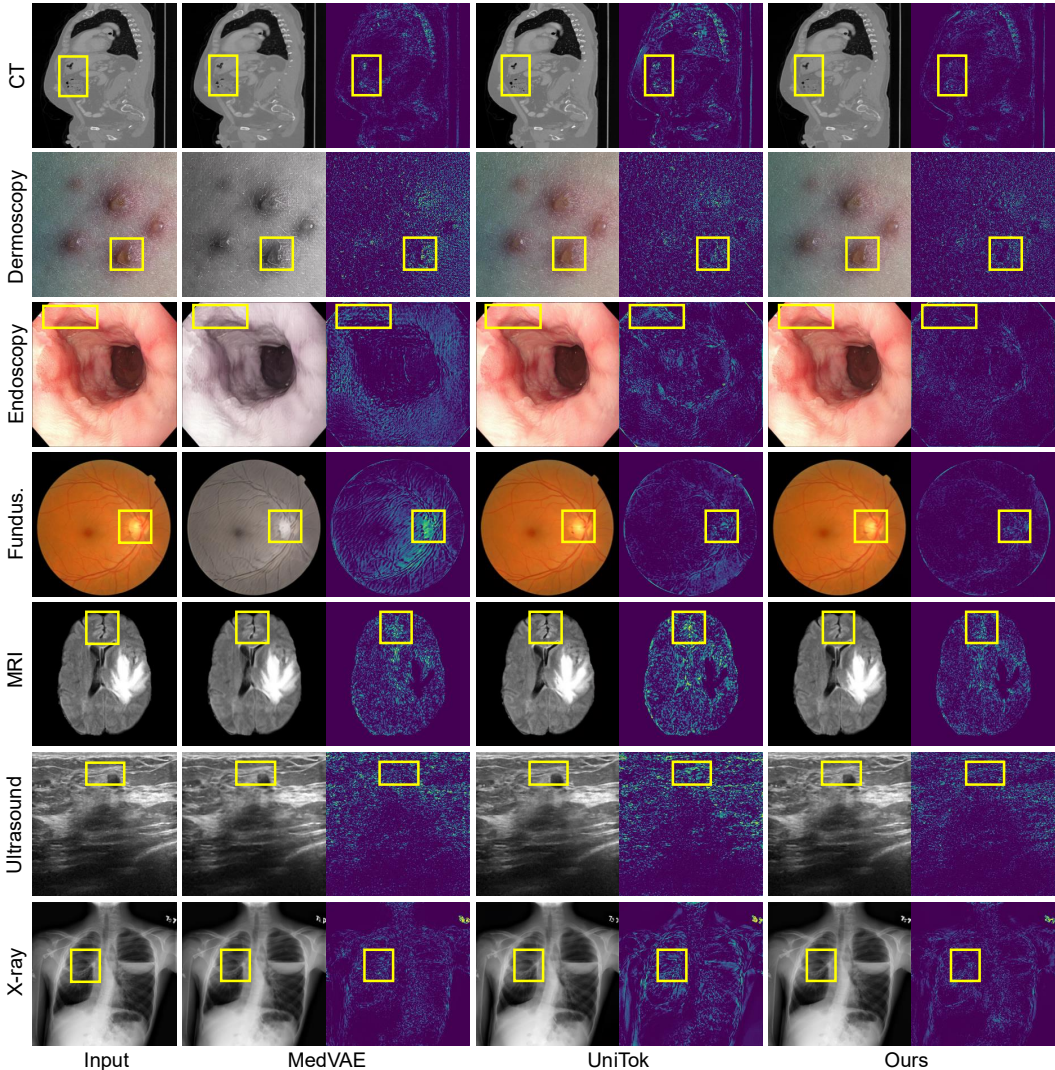


Figure S8: More reconstruction results across multiple imaging modalities. Each reconstructed image is paired with an absolute error map against the input image. Regions of interest are highlighted via yellow bounding boxes.

Return to: [Introduction](#) | [Experiments](#) | [Appendix Contents](#)

## E DISCUSSION

This section synthesizes our design choices, positioning, limitations, and societal implications. We first justify the choice of discrete codebooks (Appendix E.1). We then situate MedITok relative to representative related works, clarifying differences in objectives, training, and latent space spaces



(Appendix E.2). Next, we discuss current limitations of MedITok, and outline concrete avenues for future work (Appendix E.3). Finally, we reflect on broader impact and responsible use (Appendix E.4).

## E.1 CHOICE OF DISCRETE CODEBOOKS

Our choice of discrete tokenization is driven by the goal of building a unified latent space that can power AR models across both image synthesis and interpretation tasks in the medical domain.

To that end, discrete tokens offer the following key advantages:

- **Leveraging advances in AR modeling:** Discrete tokenization allows the medical community to directly benefit from the broader ecosystem of discrete-sequence modeling, *e.g.*, unified training objectives, any-to-any modality transfer (Zhan et al., 2024; Chen et al., 2025b), and efficient decoding and infrastructure, which are not easily transferable to continuous tokenizers.
- **Unified latent space for visual synthesis and interpretation:** Discrete tokens act as a shared representational “language” across modalities. They support AR models that can both synthesize medical images and interpret them using a single AR backbone (Lin et al., 2025). In contrast, continuous representations (*e.g.*, VAEs, CLIP) typically lack this versatility, either being hard to decode (CLIP) or poorly aligned with semantic embeddings (VAE).
- **Seamless integration with different modalities.** Discrete visual tokens are natively compatible with discrete textual tokens, enabling direct multimodal fusion in AR models without additional heads or diffusion modules. This compatibility is critical for scaling medical AR models in the style of GPT-4o, where all modalities are treated as token sequences.

## E.2 COMPARISON WITH RELATED WORKS

We situate MedITok alongside two related works: MedVAE (Varma et al., 2025) and VF-VAE (Yao et al., 2025).

MedVAE is an effective continuous variational autoencoder (VAE) designed for efficient medical image interpretation. Our primary departure from MedVAE lies in where and how semantics are bound to the latent space. Before detailing the differences, we briefly describe the training stage of interest for MedVAE and MedITok:

- MedVAE first trains a continuous VAE, then freezes the VAE encoder and decoder and learns a lightweight projector whose output is optimized so that the BiomedCLIP image embedding of the projected latent matches the embedding of the input image via an  $\ell_2$  loss, *i.e.*,  $\ell_2(\mathcal{E}_{\text{vision}}(f(z)), \mathcal{E}_{\text{vision}}(x))$ , where  $\mathcal{E}_{\text{vision}}$  denotes the pretrained BiomedCLIP vision encoder,  $f$  is the projector,  $x$  is the input image, and  $z$  is the corresponding latent.
- MedITok utilizes  $\mathcal{L}_{\text{contrastive}}(f(z), \mathcal{E}_{\text{text}}(t))$  (or  $\mathcal{L}_{\text{contrastive}}(f(z), \mathcal{E}_{\text{vision}}(x))$ , as in the first stage), where  $\mathcal{L}_{\text{contrastive}}$  is the contrastive loss, and  $t$  denotes the caption. In either stage, the encoder and decoder of MedITok are trainable.

This clearly shows the following main differences:

1. MedVAE enforces the latent  $z$  to be *perceptually close* to the input image  $x$  under BiomedCLIP, which focuses more on improving the reconstruction fidelity, while MedITok aligns  $z$  to the embedding space of BiomedCLIP so the MedITok *encodes more clinical semantics*.
2. MedVAE keeps the VAE encoder and decoder frozen in the second stage, which can be viewed as treating semantics as post-hoc *extraction* from a fixed latent. In contrast, MedITok *injects* semantics into a discrete token space since the encoder and decoder of the tokenizer is *both trainable*.
3. Since MedVAE focuses more on preserving structural details in radiological images, it did not utilize caption data for training and did not provide unified latent space for a wide range of downstream modalities and tasks. In contrast, by aligning latent tokens to BiomedCLIP embedding space, MedITok provides richer, fine-grained clinical semantics, which can be reflected in Table 2, where MedITok shows significantly better performance than MedVAE on image classification tasks.

Another related work is VF-VAE (Yao et al., 2025), which targets the reconstruction-generation trade-off in continuous VAE tokenizers for natural-image latent diffusion, proposing a single-stage joint reconstruction and alignment objective that aligns latents to a frozen vision foundation model to improve generative quality and training efficiency. The differences are as follows:

1. Primary task. VF-VAE focuses on improved visual generation using semantic constraint in latent diffusion, whereas our work targets unified generation and interpretation (e.g., classification/VQA) across diverse medical modalities. This dual-use requirement drives our design choices.
2. Methodology design. VF-VAE employs a single-stage objective to refine the latent space for better visual generation. In contrast, we use a two-stage curriculum to reach the unified goal while exploiting abundant unlabeled medical data. Moreover, VF-VAE uses cosine similarity as the alignment objective. However, as shown in Rows (x) and (xi) of Table S1, such objective significantly degrades the medical image classification performance.
3. Architecture. VF-VAE operates in a continuous VAE/diffusion setting; MedITok produces discrete, AR-ready tokens. Architecture is not the crux here, but this helps explain downstream usage differences.
4. Community. VF-VAE contributes greatly to the field of general visual generation at designing effective VAEs. Our goal, however, is to democratize a foundation visual tokenizer for medical images to serve downstream applications, with effectiveness, scalability, and general usability for the medical image community.

### E.3 LIMITATION AND FUTURE DIRECTIONS

While MedITok demonstrates strong performance across multiple medical vision tasks, there remain important considerations and limitations that motivate future work.

*First*, our two-stage training framework effectively balances structural fidelity and semantic alignment. However, optimizing simultaneously for both properties remains non-trivial. It is interesting and valuable to explore disentangling structural and semantic objectives during training (Qu et al., 2024) or jointly optimizing the tokenizer with a downstream model that unifies visual generation and interpretation (Wang et al., 2025). We opt for the current two-stage design for its simplicity and effectiveness.

*Second*, although the current version of MedITok is designed mainly for 2D medical images across multiple imaging modalities, we have also shown that MedITok can be easily adapted to 3D medical tasks that require volume processing (Table S2). Nonetheless, MedITok could benefit from future advancement such as 3D native training or mixed training using 2D images and 3D volumes, as well as evaluation on more sophisticated tasks.

*Third*, due to resource constraints, our current experiments utilize 2.4 million image-caption pairs – modest in scale compared to billion-scale training regimes in the general domain (Ma et al., 2025b). We believe that the proposed framework is scalable and can benefit significantly from larger and more diverse image-text corpora. Future efforts may explore combining public data with institution-curated pairs.

In summary, while MedITok sets a strong foundation for unified medical visual tokenization, ongoing work is needed to address the above limitations. We envision that MedITok’s flexible and expressive design can be extended to diverse downstream tasks. More broadly, we hope this work paves the way toward building scalable, general-purpose generative models that can advance medical AI and ultimately contribute to improving human health.

### E.4 BROADER IMPACT

This work presents a unified visual tokenizer tailored for medical images, offering a flexible and generalizable foundation for a wide range of medical AI applications. MedITok has the potential to accelerate the development of general-purpose medical AI systems and reduce task-specific engineering efforts. Its modular and pretrained nature also lowers the barrier for medical researchers to develop high-performance models with limited data and compute.

2700 However, this progress also raises societal considerations. Insufficient training data may lead to  
2701 biased models that underperform in underrepresented populations or clinical contexts. Additionally,  
2702 the deployment of powerful downstream generative models in medicine, based on our MedITok,  
2703 must be guided by strict ethical oversight to prevent misuse, misinformation, or over-reliance without  
2704 clinical validation. We advocate for responsible development and interdisciplinary collaboration to  
2705 ensure that such technologies benefit patients and healthcare systems.  
2706  
2707  
2708  
2709  
2710  
2711  
2712  
2713  
2714  
2715  
2716  
2717  
2718  
2719  
2720  
2721  
2722  
2723  
2724  
2725  
2726  
2727  
2728  
2729  
2730  
2731  
2732  
2733  
2734  
2735  
2736  
2737  
2738  
2739  
2740  
2741  
2742  
2743  
2744  
2745  
2746  
2747  
2748  
2749  
2750  
2751  
2752  
2753

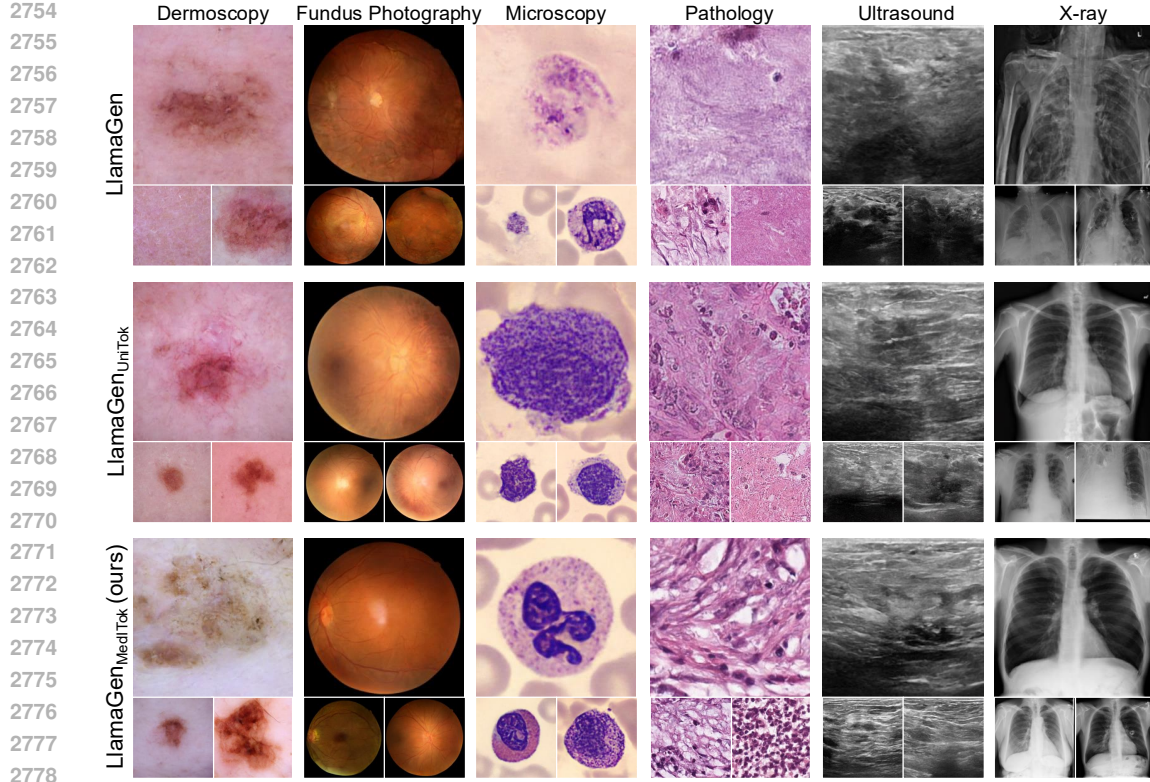
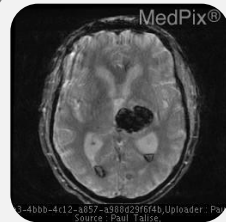


Figure S9: Qualitative comparison of medical images generated by LlamaGen models based on different visual tokenizers.



**Question:** What structures are located near the hemorrhage shown?

**Golden Answer:** Left thalamus and basal ganglia.

**LLaVA-Med:** The hemorrhage is located near the left thalamus and the left internal capsule.

**LLaVA-Med<sub>UniTok</sub>:** The hemorrhage is located near the right cerebral peduncle, which is a part of the brainstem. This area is crucial for motor and sensory pathways, as well as autonomic functions. The presence of a hemorrhage in this region can potentially affect these critical functions.

**LLaVA-Med<sub>MedITOK</sub> (ours):** The hemorrhage is located near the left thalamus and the left basal ganglia. These are critical structures in the brain involved in various functions including motor control, sensory perception, and cognitive processes. The proximity of the hemorrhage to these areas could potentially impact these functions depending on the extent and severity of the bleeding.

Figure S10: Visual question answering results on an MRI image.





**Question:** In which lobe do you see an abnormal mass in the above images?

**Golden Answer:** Right upper lobe.

**LLaVA-Med:** The abnormal mass is located in the right upper lobe of the lung.

**LLaVA-Med<sub>UniTok</sub>:** The abnormal mass is located in the right lower lobe of the lung.

**LLaVA-Med<sub>MedITok</sub> (ours):** The abnormal mass is located in the right upper lobe of the lung, as seen in the provided chest radiographs. This mass is characterized by its irregular shape and increased density compared to the surrounding lung tissue. The presence of such a mass could indicate a variety of conditions, ranging from benign tumors to malignant growths, necessitating further diagnostic evaluation such as a CT scan or biopsy to determine the exact nature and extent of the pathology.

Figure S11: Visual question answering results on a chest X-ray image.



**Question:** 图片中体积最大的器官是什么?

**Golden Answer:** 肝脏.

**LLaVA-Med:** The body with the largest volume in the image is the liver.

**LLaVA-Med<sub>UniTok</sub>:** 根据提供的CT扫描图像, 体积最大的器官是脾脏。图像显示, 脾脏的体积约为1100毫米的立方体积, 远大于其他器官的体积。脾脏的体积增大可能与某些疾病或病理状况有关, 如脾脏肿大症、脾脏炎症或脾脏肿瘤等。此外, 图像还显示肝脏、肾脏、脊柱和腹腔内的脂肪等其他器官的体积和形态。

**LLaVA-Med<sub>MedITok</sub> (ours):** 根据图像, 体积最大的器官是肝脏。肝脏在这张CT扫描中呈现出较为均匀的密度, 表明其内部结构相对均匀。

Figure S12: Visual question answering results on a CT image.

Table S5: Medical Image Datasets in Stage 1 (Part 1 of 4).

Dataset Name	Count	Dataset Name	Count
Rсна-Str-Pulmonary-Embolism-Detection (Anouk Stein et al., 2020)	5,604,627	Bcnb-Task5 (Xu et al., 2021)	76,559
Endovis2023-Surgtoolloc (Zia et al., 2025)	3,710,685	Bcnb-Task1-0 (Xu et al., 2021)	76,558
Panda (Bulten et al., 2022)	1,616,913	Bcnb-Task6 (Xu et al., 2021)	76,558
Mela (Song et al., 2022a;b;c;d)	1,403,843	Msd-Liver (Antonelli et al., 2022)	76,395
Ixi (IXI)	924,870	Ct-Org (Rister et al., 2020)	76,195
Ribfrac2020 (Jin et al., 2020)	810,265	Endovis-2021-Petraw (Huaulmé et al., 2023)	75,718
Radimagenet (Mei et al., 2022)	779,768	Head-Neck-Pet-Ct (Vallieres et al., 2017)	75,109
Autopet (Gatidis et al., 2022)	590,785	Ctspine1K (Deng et al., 2021)	72,835
Brats2023-Gli (Bakas et al., 2017c)	513,263	Bcnb-Task1-2 (Xu et al., 2021)	71,991
Atm2022 (Zhang et al., 2023a)	501,147	Lndb (Pedrosa et al., 2019)	70,292
Lidc-Idri-All-Ct (Armato III et al., 2011)	474,076	Cptac-Hnscc (National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC), 2018)	69,731
Luna16 (Setio et al., 2017)	431,694	Lung-Pet-Ct-Dx (Li et al., 2020)	66,564
Brats2023-Men (LaBella et al., 2023)	384,425	Anti-Pd-1-Melanoma (Patnana et al., 2019)	65,411
Mimic-Cxr (Johnson et al., 2019)	377,110	Nsclc-Cetuximab (Movsas et al., 2016)	64,730
Qin-Headneck (Beichel et al., 2015)	307,946	Anode09 (Van Ginneken et al., 2010)	63,250
Biomedica (Lozano et al., 2025)	291,155	Opc-Radiomics	62,726
Flare22 (Ma et al., 2024)	280,531	Acrin-Nsclc-Fdg-Pet (Kinahan et al., 2019)	62,701
Braintumour (Bakas et al., 2018)	263,310	Sln-Breast (Campanella et al., 2019)	61,968
Chexpertplus (Chambon et al., 2024)	223,460	Bcnb-Task2 (Xu et al., 2021)	61,828
Totalsegmentator-Dataset (Wasserthal et al., 2023)	218,477	Msd-Lung (Antonelli et al., 2022)	61,117
Pediatric-Ct-Seg (Jordan et al., 2022; 2021)	204,602	Bcnb-Task1-3 (Xu et al., 2021)	59,521
Acrin6668 (Machtay et al., 2013)	188,098	Midrc-Ricord-1B (Tsai et al., 2021)	59,247
Covid-19-Ny-Sbu (Saltz et al., 2021)	185,668	Bcnb-Task1-4 (Xu et al., 2021)	59,091
Bracs (Brancati et al., 2022)	177,712	Learn2Reg2022-L2R-Task1-Oasis (Hering et al., 2022b)	57,984
Abdomenct1K (Ma et al., 2021a)	172,963	Amos2022 (Ji et al., 2022)	56,217
Bone-Marrow-Cytomorphology (Matek et al., 2021)	171,378	Learn2Reg22-L2R-Oasis (Hering et al., 2022b)	52,992
Ctpelvic1K (Liu et al., 2021b)	127,315	Cataract101 (Schoeffmann et al., 2018)	52,676
Parse22 (Luo et al., 2023a)	122,629	Brats2023-Ped (Kazerooni et al., 2023)	51,769
Nih-Chest-X-Rays (Wang et al., 2017a)	112,115	Vestibular-Schwannoma-Seg (Shapey et al., 2021)	51,575
Lits (Bilic et al., 2023)	107,056	Midrc-Ricord-1A (Tsai et al., 2021)	50,913
Hnscc (Grossberg et al., 2018; 2020)	101,861	Lc25000 (Borkowski et al., 2019)	50,000
Airogs (de Vente et al., 2024)	101,280	Cptac-Luad (National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC), 2018)	48,952
Head-Neck-Cetuximab (Bosch et al., 2015)	100,356	Ct-Covid-19-August2020 (Harmon et al., 2020)	48,791
Brats2023-Met (Moawad et al., 2023)	93,775	Fastpet-Ld (Green et al., 2019)	48,097
Acrin-Flt-Breast (Kinahan et al., 2017)	91,948	Oasis2 (Marcus et al., 2010)	47,744
Bcnb-Task4 (Xu et al., 2021)	89,894	Osic-Pul-Fib-Pro (Shahin et al., 2020)	46,014
Covidx-Cxr-4 (Wu et al., 2023b)	84,802	Anti-Pd-1-Lung (Madhavi et al., 2019)	45,497
Nlst (Team, 2011)	79,194	Tcga-Luad (Albertina et al., 2016)	45,049
Cad-Pe (González et al., 2020)	78,583	Isic2020 (Rotemberg et al., 2021)	44,106
Bcnb-Task3 (Xu et al., 2021)	76,559	Longitudinal-multiple-sclerosis-lesion-segmentation (Carass et al., 2017)	41,984

Table S6: Medical Image Datasets in Stage 1 (Part 2 of 4).

Dataset Name	Count	Dataset Name	Count
Covid-19-Ar (Desai et al., 2020)	41,664	Lysto (Jiao et al., 2024)	19,990
Glis-Rt (Shusharina & Bortfeld, 2021)	41,143	Cas2023 (Chen et al., 2023)	19,200
Mura (Rajpurkar et al., 2017)	39,939	Tcga-Ov (Holback et al., 2016)	19,077
Spie-Aapm (Armato III et al., 2015)	39,670	Sicapv2 (Silva-Rodríguez et al., 2020)	18,783
Tcga-Lusc (Kirk et al., 2016b)	38,998	Vin-Big-Data (Nguyen et al., 2020)	17,999
Atlas-2 (Liew et al., 2022)	38,400	Wmh (Kuijff et al., 2019)	16,896
Spie-Aapm-Lung-Ct-Challenge (Armato III et al., 2015)	38,373	Fizpatrick17K (Groh et al., 2021; 2022)	16,577
M2Cai16-Tool (Jin et al., 2018)	37,314	Chest-Image-Pneum (Zawacki et al., 2019)	15,251
Hyperkvasir (Borgli et al., 2020)	36,329	C-Nmc-2019 (Mourya et al., 2019)	15,105
Brats-Tcga-Gbm (Bakas et al., 2017b)	35,770	Covid-19-20 (Roth et al., 2022)	15,045
Lld-Mmri2023 (Lou et al., 2023)	35,751	Aod-14800 (Abuev, 2021)	14,805
Diabetic (Platform, 2023)	35,059	Aapm-Rt-Mac (Cardenas et al., 2019)	14,080
Eyepacs (Dugas et al., 2015)	35,059	Mindboggle (Klein et al., 2017)	12,575
Ranzcr-Clip (Seah et al., 2020)	33,664	Siim-Acr-Pneumothorax (Zawacki et al., 2019)	12,053
Isic2019 (Codella et al., 2018a)	33,541	Chest-X-Ray-Images-With-Pneumothorax-Masks (Zawacki et al., 2019)	12,047
Verse20 (Sekuboyina et al., 2021b)	32,944	Han-Seg (Podobnik et al., 2023)	11,939
Covidxcxr-2 (Wang et al., 2020)	31,238	Valdo-Task1 (Sudre et al., 2024)	11,915
Lola11 (van Ginneken, 2021)	30,207	Valdo-Task3 (Sudre et al., 2024)	11,915
Rсна-Pdc (Anouk Stein et al., 2018)	29,684	Cptac-Ucec (National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CP-TAC), 2019a)	11,595
C4Kc-Kits (Heller et al., 2019)	28,843	Tcga-Stad (Lucchesi & Aredes, 2016)	11,204
Word (Luo et al., 2022)	27,154	Ultrasound-Nerve-Segmentation (Montoya et al., 2016)	11,143
Acrin-Hnscc-Fdg-Pet-Ct (Kinahan et al., 2020)	27,117	Msseg08 (Styner et al., 2008)	10,965
Kits2021 (Heller et al., 2020)	26,503	Wsss4Luad (Han et al., 2022)	10,091
Exact09 (Lo et al., 2012)	25,560	Medfm-Colon-2023 (Wang et al., 2023)	10,009
Bcnb-Task1-1 (Xu et al., 2021)	25,370	Knee-Osteoarthritis-Dataset (Chen, 2018)	9,766
Survisdom (Zia et al., 2021)	24,360	Segthor (Lambert et al., 2020)	9,661
Brats-Tcga-Lgg (Bakas et al., 2017a)	23,336	Brain-Ptm (Avital et al., 2019; Nelkenbaum et al., 2020)	9,600
Tcga-Ucec (Erickson et al., 2016)	22,946	Msd-Colon (Antonelli et al., 2022)	9,191
Tcga-Kirc (Akin et al., 2016)	22,644	Covid19Ctscans (Jun et al., 2020)	9,119
Cptac-Sar (National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CP-TAC), 2019b)	22,432	Cholect50 (Nwoye et al., 2023)	8,919
Crossmoda2023 (Dorent et al., 2023)	21,981	Msd-Pancreas (Antonelli et al., 2022)	8,666
Cptac-Cm (National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CP-TAC), 2018b)	21,867	Fumpe (Masoudi et al., 2018)	8,402
Brats2023-Ssa (Adewole et al., 2023)	20,910	Lctsc (Yang et al., 2017)	8,300
Pancreas-Ct (Roth et al., 2015)	20,709	Ct-Vs-Pet-Ventilation-Imaging (Eslick et al., 2018)	8,252
Vessel2012 (Rudyanto et al., 2014)	20,442	Head-Neck-Radiomics-Hn1 (Aerts et al., 2014)	8,161
Yangxi (Liu et al., 2019)	20,394	Qin-Breast (Li et al., 2015)	8,051
Msseg2016 (Commowick et al., 2018)	20,352	Chaos-Task-4 (Kavur et al., 2021)	7,977
Oia-ODir (Peking University International Competition on Ocular Disease Intelligent Recognition (ODIR-2019), 2019)	19,992	Pannuke (Gamper et al., 2019; 2020)	7,810

Table S7: Medical Image Datasets in Stage 1 (Part 3 of 4).

Dataset Name	Count	Dataset Name	Count
Sppin2023 (Buser et al., 2025)	7,616	Pad-Ufes-20 (Pacheco et al., 2020)	2,298
Atlas2023 (Quinton et al., 2023)	7,364	Msd-Spleen (Antonelli et al., 2022)	2,169
Msd-Hepaticvessel (Antonelli et al., 2022)	6,859	Breakhis-100X (Spanhol et al., 2015)	2,081
Mmwhs (Zhuang, 2018)	6,400	Breakhis-200X (Spanhol et al., 2015)	2,011
Hsa-Nrl (Zhu et al., 2021)	6,160	Breakhis-40X (Spanhol et al., 2015)	1,991
Coronahack (Praveen Govi, 2019)	5,933	Breakhis-400X (Spanhol et al., 2015)	1,820
Rus-Chn (Baidu AI Studio, 2021)	5,921	Cptac-Pda (Consortium et al., 2018)	1,792
Dhrf (Derbi Hackathon Organizers, 2022)	5,680	Tiger-Wsirois-Roi-Level-Tissue-Cells (van Rijthoven et al., 2022)	1,775
Aptos2019-Blindness-Detection (apt)	5,590	Breast-Diagnosis (Wolberg et al., 1995)	1,656
Curious2019 (Xiao et al., 2019)	5,376	Cmb-Gec (Biobank, 2022a)	1,625
Cmb-Mel (Biobank, 2022b)	5,289	Riga-Dataset (Almazroa et al., 2018)	1,617
Clust15-2D (Luca et al., 2018)	5,255	Refuge2-Clis (Fang et al., 2022)	1,600
Cmmd (Cui et al., 2021)	5,202	Harvardglaucoma-1547 (Kim, 2018)	1,544
Tcga-Hnsc (Zuley et al., 2016)	5,172	Tcga-Kich (Linehan et al., 2016)	1,484
Continuous-Registration-Task3 (Baheti et al., 2021)	5,120	Papilledema (pap, 2020)	1,369
Messeg (Commowick et al., 2018)	5,120	Continuous-Registration-Task6 (Hering et al., 2022a)	1,280
Node21 (Sogancioglu et al., 2024)	4,882	Isbi2016-Part3 (Gutman et al., 2016)	1,279
Conic2022 (et al., 2021)	4,870	Isic2016-Task1 (Gutman et al., 2016)	1,279
Lag-4854 (Li et al., 2019)	4,854	Fusc2021 (Wang et al., 2024a)	1,210
Medfm-Chestdr-2023 (OpenMEDLab, 2023)	4,848	Hvsmr-2016 (Pace et al., 2015)	1,152
Stageii-Colorectal-Ct (Tong & Li, 2022)	4,672	Osteosarcoma-Tumor-Assessment (Leavey et al., 2019)	1,143
Naf-Prostate (Kurdziel, 2015)	4,664	Isic2016-Task2B-Globules (Gutman et al., 2016)	1,142
Chest-X-Ray-Pa (Asraf & Islam, 2021)	4,574	Isic2016-Task2B-Streaks (Gutman et al., 2016)	1,142
Lungct-Diagnosis (Grove et al., 2015)	4,155	Jsiec (Cen et al., 2021)	997
Covid19-Radio-Data (Chowdhury et al., 2020)	3,886	Isles2022 (Hernandez Petzsche et al., 2022)	938
Structseg2019-Subtask1 (Organizers, 2019)	3,634	Covid-19-Ct-Cxr-Det (Peng et al., 2020)	929
Structseg2019-Subtask4 (Organizers, 2019)	3,634	Covid-19-Ct-Cxr (Peng et al., 2020)	918
Structseg2019-Subtask2 (Organizers, 2019)	3,413	E-Ophta (Decenciere et al., 2013)	905
Qin-Lung-Ct (Kalpathy-Cramer et al., 2015)	3,586	Dao-Slocpasa (Chiu et al., 2013)	840
Structseg2019-Subtask3 (Organizers, 2019)	3,413	Continuous-Registration-Task5 (Klein et al., 2009)	813
Tcga-Coad (Network et al., 2012)	3,093	Fives (Jin et al., 2022)	800
Tcga-Prad (Abeshouse et al., 2015)	3,007	Segpc2021 (Gupta et al., 2023)	773
Bidr-2838 (Islam et al., 2021)	2,838	Paraguay-757 (Benítez et al., 2021)	757
Refuge2 (Fang et al., 2022)	2,800	Mudi2019 (Pizzolato et al., 2020)	695
Cptac-Ccrcc (National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CP-TAC), 2018a)	2,798	Pulmonary-Chest-X-Ray-China (Jaeger et al., 2014a; Candemir et al., 2014b)	662
Isic2017 (Codella et al., 2018b)	2,748	Glaucoma-Detection (Shikamaru, 2021)	650
Verse19 (Sekuboyina et al., 2021a)	2,650	Beh-634 (Islam et al., 2022)	634
Palm19 (Fang et al., 2024)	2,379		



Table S8: Medical Image Datasets in Stage 1 (Part 4 of 4).

Dataset Name	Count	Dataset Name	Count
Retina-Cataract-Dataset (yiweichen04, 2016)	601	Orvs (Sarhan et al., 2021)	202
Idrid (Porwal et al., 2020)	597	Gamma3 (Wu et al., 2023a)	200
Sz-Cxr (Stirenko et al., 2018)	566	Fund-179 (Yin et al., 2013)	179
Cmb-Pca (Fedorov et al., 2023)	532	Drac2022-Taska2 (Qian et al., 2023)	174
Crass (Hogeweg et al., 2012)	518	Drac2022-Taska3 (Qian et al., 2023)	174
Herlev (Jantzen et al., 2005)	504	Tcga-Read (Kirk et al., 2016a)	168
Papila (Kovalyk et al., 2022)	488	Glas (Sirinukunwattana et al., 2017)	165
Rimonedl (Batista et al., 2020)	485	Drac2022-Taska1 (Qian et al., 2023)	151
Fetoscopy-Placenta-Dataset (Bano et al., 2020)	482	Tiger-Wsirois-Roi-Level-Tissue-Bcss (Amgad et al., 2019)	151
Tcga-Blca (Kirk et al., 2016a)	439	Tcga-Lgg (Kirk et al., 2016a)	145
Drimdb (Prentašić et al., 2013)	428	Pulmonary-Chest-X-Ray-Montgomery (Jaeger et al., 2014b; Candemir et al., 2014a)	138
Toxofundus (Cardozo et al., 2023; Alam et al., 2023)	411	Bcss (Amgad et al., 2019)	121
Adam (Timmins et al., 2021)	400	Drishti-Gs-Cup (Sivaswamy et al., 2014)	101
Ph2 (Mendonça et al., 2015)	400	Drishti-Gs-Od (Sivaswamy et al., 2014)	101
Crown (Vos et al., 2024)	384	Avn (Nguyen et al., 2013)	90
Rose (Ma et al., 2021b)	348	Jsr-Lung (Shiraishi et al., 2000)	60
Mias (Pisano & Yaffe, 2005)	322	Breast-Cancer-Cell-Seg (Gelasca et al., 2008)	58
Covid-19-Image-Dataset (Sohan, 2020)	317	Monuseg (Kumar et al., 2020)	51
Gamma (Wu et al., 2023a)	300	Hrf (Budai et al., 2013)	45
Monusac20 (Verma et al., 2021)	283	Drhagis (Holm et al., 2017)	40
Rod (Grace Maria Binu, 2023)	281	Drive (Staal et al., 2004)	40
Jsr (Shiraishi et al., 2000)	247	Rite (Hu et al., 2013)	40
Jsr-Gender-Cls (Shiraishi et al., 2000)	247	Hrf-Quality-Cls (Budai et al., 2013)	36
Tcga-Sarc (Kirk et al., 2016a)	241	Retinacheck (Dashtbozorg et al., 2016)	30
Crag (Graham et al., 2019a)	213	Olives-Fundus-Photography (Prabhushankar et al., 2022)	14
Panda-Radboud (Nir et al., 2018a)	206	Occmcpv (Chen et al., 2024a)	8

Table S9: Medical Image Datasets in Stage 2.

Dataset Name	Count	Dataset Name	Count
Biomedica (Lozano et al., 2025)	1,216,529	Mimic-Cxr (Johnson et al., 2019)	107,684
Gmai-VI-5.5M (Li et al., 2024)	671,824	Rocov2 (Rückert et al., 2024)	59,212
Medicat (Subramanian et al., 2020)	204,772	Pmc-Oa (Lin et al., 2023)	36,386
Llava-Med-Instruct-Fig-Captions (Li et al., 2023)	122,843	Mm-Retinal (Wu et al., 2024)	3,577

Table S10: Medical Image Datasets for Image Reconstruction Evaluation.

Dataset Name	Count	Dataset Name	Count
Ivygap-Radiomics (Pati et al., 2020)	8,456	Monkeypox (Ali et al., 2022)	802
Chestx-Det (Lian et al., 2021)	3,578	Breast-Ultrasound-Images-Dataset (Al-Dhabyani et al., 2020)	647
Aapm-lowdose-ct (McCollough et al., 2017)	3,413	Ddti (Pedraza et al., 2015)	637
Btvc-Cervix (Landman et al., 2015)	3,039	Hie2023 (Bao et al., 2025)	554
Surgt (Cartucho et al., 2024)	2,933	Digestpath19-Clis (Da et al., 2022)	455
Silver07 (Heimann et al., 2009)	2,291	EndoCV2020-EDD (Ali et al., 2020)	386
Derm7Pt (Kawahara et al., 2018)	2,013	Mednode (Giotis et al., 2015)	170
Messidor (Decenci�re et al., 2014)	1,748	Gleason (Nir et al., 2018b)	103
Rсна-Bone-Age (Halabi et al., 2019)	1,596	Consep (Graham et al., 2019b)	41
Hmc-Qu (Kiranyaz et al., 2020)	1,269	Chase (Fraz et al., 2012)	28
Covidgr (Tabik et al., 2020)	852	Stare (Hoover et al., 2000)	20

Table S11: Downstream Medical Vision Tasks Datasets. “CLS”: classification. “M2I”: modality-to-image synthesis. “VQA”: visual question answering.

Dataset	Train	Test	Modality	Task Type	Classes
PneumoniaMNIST (Kermany et al., 2018)	4,708	1,148	X-ray	CLS	2
PathMNIST (Kather et al., 2019)	89,996	500	pathology	CLS; M2I	9
ChestMNIST (Wang et al., 2017b)	78,468	500	X-ray	M2I	14
BloodMNIST (Acevedo et al., 2020)	11,959	500	microscopy	M2I	8
DermaMNIST (Tschandl et al., 2018; Codella et al., 2019)	7,007	500	dermoscopy	CLS; M2I	7
RetinaMNIST (Liu et al., 2022)	1,080	500	fundus photography	CLS; M2I	5
BreastMNIST (Al-Dhabyani et al., 2020)	546	234	ultrasound	CLS; M2I	2
Pubmed-Vision-Caption (Chen et al., 2024c)	555,103	0	Unknown	VQA	–
Pubmed-Vision-VQA (Chen et al., 2024c)	100,000	0	Unknown	VQA	–
VQARAD-Test (Lau et al., 2018)	0	451	Unknown	VQA	–
Slake-Test (Liu et al., 2021a)	0	2,094	Unknown	VQA	–
Slake-Val (Liu et al., 2021a)	0	2,099	Unknown	VQA	–