Bridging the Dimensionality Gap: A Scoping Review of 3D Vision-Language Models in Medical Imaging

Anonymous ACL submission

Abstract

While large multimodal models have achieved remarkable success in general domains, adapting them to medical imaging faces a fundamental dimensionality gap: transition from 2D snapshots to 3D volumetric data (e.g., CT, MRI). This review systematically examines 37 studies on 3D Vision-Language Models in healthcare, capturing the rapid research surge of 2024-2025 in this emerging field. We further propose a categorical framework that classifies these studies by multimodal fusion mechanisms (projection, attention, and adapters), volumetric encoding strategies (slice-based vs. native 3D), and language processing (encoder vs. foundation model). Our analysis highlights the growing adoption of parameter-efficient fine-tuning due to computational constraints, but significant challenges remain, including hallucinations, a lack of spatial grounding, and misalignment between evaluation metrics and clinical utility. This survey aims to clarify current methodologies and identify future directions in volumetric medical AI.

1 Introduction

The integration of Computer Vision (CV) and Natural Language Processing (NLP) has fundamentally reshaped artificial intelligence research, enabling models to jointly reason over visual and language modalities (Karpathy and Fei-Fei, 2015; Lu et al., 2019; Tan and Bansal, 2019; Radford et al., 2021). General-domain Foundation Models (FMs) and Vision-Language Models (VLMs), such as CLIP (Radford et al., 2021) and GPT-4V (OpenAI et al., 2023), have demonstrated exceptional performance in cross-modal representation learning and generation. Despite this progress, a critical bottleneck limits their adoption in clinical settings: the dimensionality mismatch between general-purpose VLMs and medical imaging data. Most existing VLMs are designed for 2D images (Lu et al., 2019; Tan and Bansal, 2019; Radford et al., 2021),

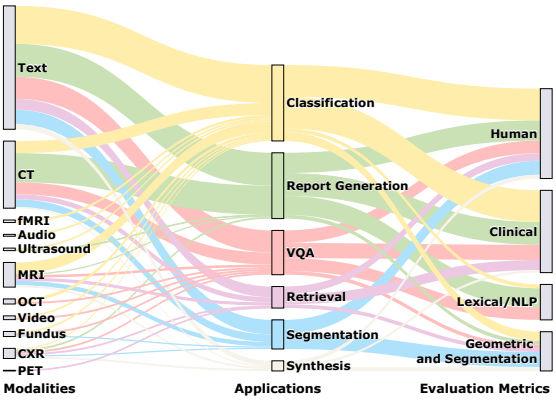

Figure 1: The relationships between data modality, downstream vision-language task, and evaluation metrics examined in this review.

whereas key diagnostic modalities, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), produce three-dimensional volumes (Milletari et al., 2016). Modeling volumetric data exponentially increases computational complexity and requires preserving spatial and depth information, which is often lost in 2D pre-trained encoders (Yang et al., 2021; Zhang et al., 2021). Consequently, simply extending 2D architectures to 3D medical imaging often yields degraded performance or unjustifiable computational costs.

In response to this challenge, recent research activity has rapidly expanded around 3D Medical VLMs (hereafter 3D-MedVLMs) (Lu et al., 2025; Chang et al., 2025; Xin et al., 2025) (Figure 1). Proposed solutions range from decomposing 3D volumes into pseudo-2D sequences to developing native volumetric encoders and utilizing Large Language Models (LLMs) via lightweight adaptation mechanisms (Dack et al., 2023; Kakkar et al., 2024; Liu et al., 2025; Luo et al., 2025). However, these efforts remain fragmented across design choices, such as different feature fusion, pre-training objectives, and evaluation metrics, making systematic comparison difficult.

Unlike general or 2D-focused surveys (Caffagni et al., 2024; Sim et al., 2025), this review focuses on the unique challenges of 3D-MedVLMs. Furthermore, we distinguish our work from the clinically-oriented roadmap of Wu et al. (2025) by identifying a broader set of studies and by focusing on the unique methodological and practical challenges posed by volumetric data in clinical settings.

Our contributions are threefold: (i) Systematic Classification: We provide a framework to categorize 3D-MedVLMs based on vision encoding and multimodal fusion techniques (Figure 4). (ii) Reasoning Analysis: We analyze how recent models capture intrinsic medical dependencies. (iii) Clinical Alignment: We identify persistent challenges in clinical evaluation, highlighting the discrepancy between standard NLP metrics and clinical factual correctness for 3D-MedVLMs.

2 Scope and Literature Selection

Our scoping review adheres to the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (Figure 2) (Tricco et al., 2018). We searched 4 databases: PubMed[1], IEEE Xplore[2], ACM Digital Library[3], and ACL Anthology[4], covering studies published from 2021 to 2025. Our search strategy was designed to systematically identify studies at the intersection of volumetric imaging and multimodal learning. We employed Boolean queries that combined three primary concept groups: (i) 3D imaging modalities, (ii) multimodal architectures, and (iii) clinical text or downstream tasks. The complete list of database-specific search queries is detailed in Appendix A.

We then screened the articles based on their titles and abstracts according to predefined criteria. Inclusion criteria required studies (i) were published in English, (ii) underwent peer review, and (iii) presented and benchmarked a model. Exclusion criteria filtered out (i) studies unrelated to 3D-MedVLMs, (ii) non-English publications, and (iii) secondary literature such as systematic reviews, case studies, or descriptive studies lacking experimental results. After screening, we extracted metadata from each included study, such as models, datasets, applications, results, and reported limitations. To ensure accuracy, two annotators cross-verified the study selection and metadata extraction, with a third one resolving any disagreements.

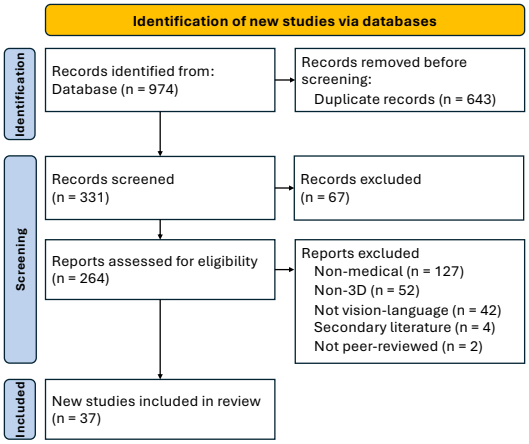

Figure 2: PRISMA-ScR flow diagram.

From an initial set of 974 papers retrieved from databases, we removed 643 duplicates. During the initial screening, 6 papers were excluded, and 3 additional papers were excluded for misaligned objectives or lack of relevance to 3D medical VLMs. Ultimately, 37 studies met the inclusion criteria and form the basis of this review. Important metadata, including resource links, architectures, training strategies, and downstream tasks for the studies, are presented in Table 1.

3 Multimodal Fusion

The core challenge in developing 3D-MedVLMs lies in aligning 3D volumetric data with textual representations. Existing approaches can be broadly categorized into three interaction paradigms, distinguished by the level at which visual and textual modalities intersect (Figure 3).

3.1 Global Alignment (Dual-Encoder)

Global alignment strategies prioritize computational efficiency by encoding visual and textual modalities in separate streams that interact only at the final embedding stage. This `Dual-Encoder` approach is well-suited to zero-shot classification and retrieval. Within this paradigm, methods differ in how cross-modal interaction is realized.

`Projection-based` alignment, popularized by CLIP (Radford et al., 2021), employs lightweight linear layers to map modalities into a shared latent space. Extensions such as RadCLIP (Lu et al., 2025) introduce slice-pooling adapters that aggregate 2D slice embeddings into a unified volumetric representation, enabling global image-text match-

[1] https://pubmed.ncbi.nlm.nih.gov/
[2] https://ieeexplore.ieee.org/
[3] https://dl.acm.org/
[4] https://aclanthology.org/

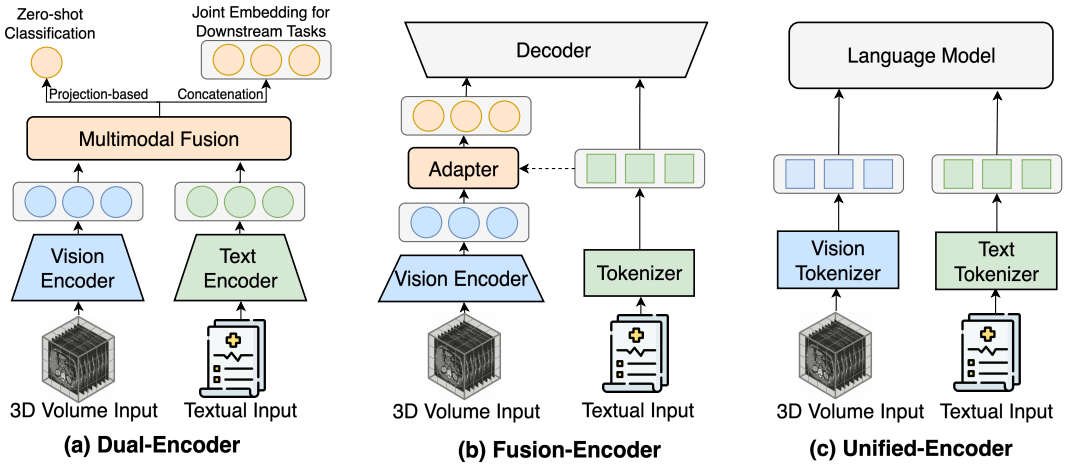

Figure 3: Overview of vision-language model architectures. This figure illustrates three architectural paradigms for studies integrating 3D medical imaging and clinical text. a, **Dual-Encoder** employs separate encoders to process visual and textual data in parallel, followed by multimodal fusion to achieve global alignment (e.g., CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021)). b, **Fusion-Encoder** utilizes a vision encoder and adapter module to connect vision embeddings with tokenized text for unified decoding (e.g., LLaVA (Liu et al., 2023), BLIP-2 (Li et al., 2023), and Flamingo (Alayrac et al., 2022)). c, **Unified-Encoder** bypasses complex fusion mechanisms by projecting both visual and textual inputs into a unified language model token space (e.g., Med3DVLM (Xin et al., 2025) and RadGenome-ChestCT (Zhang et al., 2025)).

ing. Similarly, HCL-AL (Gao et al., 2025) aligns 3D features with text through learned projections for anatomical retrieval. `Concatenation-based` fusion, on the other hand, directly concatenates features from independent encoders before a task-specific head. Zhao et al. (2025) follows this approach by joining ultrasound video features with clinical text embeddings to predict thyroid nodule invasiveness, while MultiModalGAN (Ghadekar et al., 2025) concatenates CNN image features with BERT embeddings to condition generative adversarial networks for image synthesis.

While Dual-Encoder offers computational efficiency, it often sacrifices fine-grained contextual information (Wang et al., 2022).

3.2 Dense Interaction (Fusion-Encoder)

Dense interaction paradigms address the above limitation by integrating visual and textual modalities at intermediate layers, enabling discriminative tasks such as segmentation and grounding. These `Fusion-Encoder` architectures leverage attention-based deep fusion to capture token-level interactions between visual and textual sequences.

Such a strategy is essential for tasks that require precise semantic grounding. For example, Ye et al. (2024) employs bi-attention blocks that combine self-attention over instructions with cross-attention over MRI features to drive text-prompted segmentation. Some architectures further adopt the `Encoder-Decoder` framework, pairing a fusion encoder with a shared language decoder (e.g., VL-BERT (Liu et al., 2021b)).

Despite their effectiveness in localization and grounding, Fusion-Encoder incurs higher computational overhead than Dual-Encoder (Li et al., 2022).

3.3 Generative Alignment (Unified-Encoder)

Unlike previous paradigms that coordinate separate encoders, `Generative Alignment` harnesses the unified processing capabilities of foundation models to jointly handle visual and textual modalities. These `Unified-Encoder` models project 3D visual features directly into the token space of an LLM, thereby eliminating explicit fusion mechanisms and capitalizing on pre-trained foundation models.

In this approach, the "architecture" is defined by the adapter-based mechanism that bridges modalities. For example, Med3DVLM (Xin et al., 2025) maps 3D visual features into a Qwen-based LLM (Qwen et al., 2024), enabling joint reasoning over 3D visual tokens and textual instructions. Given the high cost of full fine-tuning, frameworks like RadGenome-ChestCT (Zhang et al., 2025) employ Parameter-Efficient Fine-Tuning (PEFT) by inserting trainable bottleneck adapters into frozen Llama (Grattafiori et al., 2024) backbones, thereby

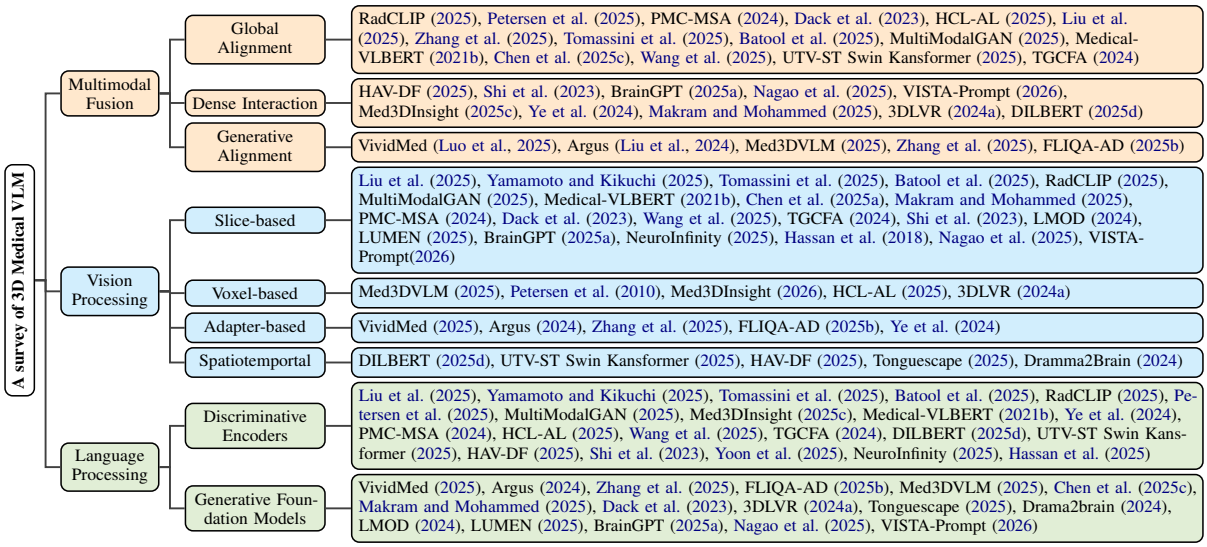

Figure 4: Categorical framework of 3D medical VLM research and representative studies included in this work.

preserving general-world knowledge in LLM while adapting to 3D medical data. Hybrid approaches such as FLIQA-AD (Chen et al., 2025b) integrate 3D adapters directly within Vision Transformers before feeding features into text generators such as FLAN-T5 (Chung et al., 2022).

While Unified-Encoder offers strong flexibility and performance, it remains susceptible to hallucination when the LLM relies on parametric knowledge rather than visual input (Li et al., 2023).

4 Single Modal Representation

While multimodal fusion enables cross-modal reasoning, the efficacy of 3D-MedVLMs relies on robust feature processing. The primary challenge lies in the `vision processing`, where models must efficiently process high-dimensional 3D data. In parallel, `language encoders` require strategies tailored to specific downstream tasks. This section examines the encoding mechanisms for them.

4.1 Vision Processing

The dimensionality mismatch between 2D models and 3D volumetric medical data (CT, MRI, PET) has led to four main encoding strategies.

Slice-based Feature Extraction decomposes volumes into 2D slices processed by 2D convolutional or Transformer pretrained on large datasets (Deng et al., 2009; Mei et al., 2022). It is computationally efficient at the cost of full-volumetric context, or when aggregating 2D slices provides sufficient contextual information (Lu et al., 2025; Kunhimon et al., 2024).

Native 3D architectures operate directly on voxel patches to capture the intricate spatial dependencies inherent in volumetric organs. This strategy is essential for tasks requiring precise anatomical localization along the Z-axis but entails high memory and computational demands. To overcome these challenges, Chen et al. (2025c); Xin et al. (2025) incorporate specialized transformer-based models to efficiently capture volumetric features.

Adapter-Based Strategies extends 2D encoders for volumetric data by combining 2D pretraining with 3D spatial awareness (Chen et al., 2025b). These approaches “inflate" pretrained 2D backbones by inserting lightweight adapters, allowing 3D input processing while retaining large-scale 2D pretraining, as seen in FLIQA-AD (Chen et al., 2025b) and VividMed (Luo et al., 2025).

Spatiotemporal Modeling treats depth or time as a sequence, and applies specialized attention or convolutional mechanisms to capture temporal continuity (Chen et al., 2025d). For example, Zhao et al. (2025) applies a Video Swin (Liu et al., 2021c) backbone to ultrasound videos to learn temporal consistencies for predicting thyroid nodule invasiveness. However, spatiotemporal modeling remains relatively unexplored in medicine.

4.2 Language Processing

Clinical text processing in 3D-MedVLMs has progressed from simple semantic-embedding extraction to complex autoregressive reasoning, reflecting a shift from static image-text alignment to instruction-following and generative capabilities. Existing studies identify two primary paradigms.

Discriminative Encoders are suitable for segmentation, and classification tasks. These encoder-only models map clinical text to a fixed latent space (Liu et al., 2025), primarily using transformer-based encoders such as BERT and CLIP (Liu et al., 2021b, 2025). Beyond free-text, Petersen et al. (2025) encodes tabular health data into natural-language descriptions, thereby enabling alignment with 3D MRI features.

Generative Foundation Models have recently been adopted to address complex language understanding and reasoning tasks, such as report generation and visual question answering (VQA). Med3DVLM (Xin et al., 2025) exemplifies this line of work by integrating Qwen2.5-7B (Qwen et al., 2024), allowing the system to process 3D visual tokens alongside textual instructions. Similarly, Argus (Liu et al., 2024), RadGenome-ChestCT (Zhang et al., 2025), and BrainGPT (Li et al., 2025a) adopt Llama models (Grattafiori et al., 2024) as their linguistic backbone, leveraging their broad real-world knowledge to generate detailed, contextually accurate radiology reports.

5 Training Strategies

Pre-training. 3D-MedVLMs's capacity to interpret clinical text is established through pretraining objectives. Two primary pretraining strategies are commonly used. (1) `Contrastive Learning` aligns the global image features with corresponding text reports. Standard CLIP uses softmax loss, which requires large batch sizes to provide sufficient negative samples. However, this is often impractical for 3D-MedVLMs due to GPU memory constraints (Li et al., 2019). To address this, approaches such as SigLIP employ a pairwise Sigmoid loss that decouples performance from batch size (Xin et al., 2025), enabling effective alignment even with small batch sizes. However, such global alignment methods may associate a medical term (e.g., "pneumonia") with broad visual patterns (e.g., general hazy texture), without accurately localizing the pathology. (2) `Masked Signal Modeling` (MSM) complements contrastive learning by learning dense, local dependencies through reconstructing masked portions of input. This encourages the encoder to model the underlying structure of 3D volumes or clinical text rather than solely on high-level semantics (Chen et al., 2025d; Ye et al., 2024). However, MSM can be computationally expensive for volumetric data and may overfit to low-level noise if not combined with a discriminative loss. Still, it provides the fine-grained spatial precision that is lost in global contrastive methods.

Fine-Tuning. As 3D-MedVLMs scales to billions of parameters, full model fine-tuning becomes computationally burdensome and prone to forgetting. `Parameter-Efficient Fine-Tuning` (PEFT) addresses this by only updating a fraction of the model's weights. `Additive methods` introduce new trainable parameters to fixed backbones. Common approaches include Adapters, which inject bottleneck layers (Chen et al., 2025b), and Head Tuning, which fine-tunes only the lightweight decision heads (Luo et al., 2025). However, these methods rely heavily on the relevance of the frozen features (Liu et al., 2021b). `Selective methods` fine-tune a subset of existing parameters. Approaches such as Medical-VLBERT (Liu et al., 2021b) employ alternate learning to selectively optimize the language decoder separate from joint multimodal updates. However, the performance of this strategy is highly sensitive to the training schedule, requiring a balanced ratio of pretraining to transferring iterations to prevent performance degradation. Finally, `Reparameterization methods` (e.g., LoRA) constrain optimization to a low-rank subspace to reduce memory consumption while maintaining fine-tuning efficiency (Liu et al., 2024). However, shared low-rank constraints often fail to capture unique intra-group heterogeneity, leading to demographic fairness disparities (Li et al., 2025b).

6 Reasoning

3D-MedVLMs require advanced reasoning capabilities. Here, we summarize how included studies capture the intrinsic dependencies of medical data.

Spatial and Topological Reasoning. Standard global embeddings often discard the anatomical hierarchy and connectivity required for advanced reasoning. Recent methods address this by explicitly encoding 3D geometry and enforcing fine-grained visual-textual alignment. Du et al. (2024a) introduces a Topological Feature Extractor based on persistent homology to capture structural invariance beyond pixel-level features. Gao et al. (2025) proposes Hierarchical Contrastive Learning to enforce semantic alignment across coarse-to-fine anatomical hierarchies and establishes a broad anatomical context (e.g., body region) before refining its focus to specific substructures. VividMed (Luo et al.,

2025) further bridges semantic understanding and localization by integrating a VLM with a localization module that utilizes the hidden states of special enclosure tokens to generate segmentation masks and bounding boxes for both 2D and 3D inputs.

Temporal and Longitudinal Reasoning. Diagnostic reasoning also requires capturing both the longitudinal progression of pathology and the spatiotemporal continuity of real-time imaging. 3DVLR (Du et al., 2024a) models "evolve anatomical landscapes" by using optimal transport theory to align feature distributions across different time stamps, thus distinguishing stable anatomical structures from changing pathological conditions (e.g., tumor growth) and enabling the generation of narratives that reason about disease progression. For dynamic imaging, Chen et al. (2025d) propose DILBERT, which aggregates visual geometric features with named entity linguistic features across video frames and employs multimodal distillation to align dynamic visual evidence with diagnostic text, ensuring consistent interpretation of real-time scans such as ultrasound.

Clinical Chain-of-Thought (CoT) and Causal Reasoning. Medical diagnosis also involves a deductive process of hypothesis generation and verification, rather than single-step classification, motivating a shift towards CoT reasoning to mirror clinical workflows. Gai et al. (2025) trains models to output intermediate reasoning steps using datasets (e.g., R-RAD, R-SLAKE) annotated with reasoning traces. Holistic Narrative Synthesis extends reasoning to 3D data by structuring visual findings, such as lesion location and size, into logically flowing narratives. For example, BrainGPT (Li et al., 2025a) produced narratives rated indistinguishable from human-written reports in 74% of cases. Making up for the lack of reasoning-focused metrics, RaTEScore (Zhao et al., 2024) assesses factual correctness of key medical entities (e.g., diagnostic outcomes, anatomical details), and FORTE (Li et al., 2025a) captures report completeness.

7 Benchmark Dataset

In the 3D medical domain, datasets have historically lagged behind 2D counterparts. Our review highlights that 3D datasets are typically smaller, less accessible, and often lack associated reports in prior studies. To address this gap, increasing efforts have focused on curating large-scale public 3D image datasets with text annotations (Zhang et al., 2025) or employing synthetic data augmentation (Ghadekar et al., 2025; Chen et al., 2025a) to increase both dataset size and diversity (Table 2).

Among these, CT-RATE (Hamamci et al., 2025a) is particularly noteworthy. With more than 50,000 volumes from over 20,000 patients, it offers a substantial scale advantage over other datasets. Additionally, text reports were authored by multiple radiologists with varied writing styles, further enhancing the dataset's robustness.

Other datasets, such as Ichinose et al. (2023) and RadGenome-ChestCT (Zhang et al., 2025), address a different need: image grounding. While many datasets provide image-text pairs, such as a slice of the lungs with the caption "nodule in both lungs", they often lack precise localization. Grounding-focused datasets address this gap, enabling models like VividMed (Luo et al., 2025) to produce both textual reports and corresponding bounding boxes or segmentation masks, thereby bridging the semantic gap between "what" and "where".

Another group of datasets moves beyond simple answers to include rationales and explanations. While traditional datasets like VQA-RAD (Lau et al., 2018) and SLAKE (Liu et al., 2021a) included short-form answers, newer initiatives such as R-RAD and R-SLAKE have annotated the intermediate reasoning steps required for medical decision-making (Gai et al., 2025). Similarly, instruction-response datasets like M3D (Bai et al., 2024) have emerged to support complex reasoning.

In contrast to these text-rich initiatives, the visual backbones often rely on anatomical understanding through segmentation masks. Leading this category is TotalSegmentator (Wasserthal et al., 2023), which provides 5,000 CT volumes with comprehensive masks for 117 anatomical classes. Similarly, AbdomenAtlas (Li et al., 2025c), a large-scale dataset comprising over 8,000 volumes, and AbdomenCT-1K (Ma et al., 2022) exemplify ongoing efforts to scale segmentation benchmarks.

8 Evaluation Metrics

Evaluating 3D-MedVLMs requires a comprehensive assessment of technical performance, clinical utility, and robustness. Accordingly, metrics should jointly capture vision-task accuracy and text quality, including fidelity to visual evidence, linguistic coherence, and clinical relevance (Liu et al., 2024; Xin et al., 2025; Zhang et al., 2025).

Geometric and Segmentation Metrics. For

anatomical segmentation and visual grounding, geometric metrics quantify spatial accuracy. Volumetric overlap between predicted and ground-truth masks is commonly measured using the Dice Similarity Coefficient (DSC) and the Intersection over Union (IoU) (Ye et al., 2024; Wang et al., 2025). Boundary accuracy is assessed using the 95th-percentile Hausdorff Distance (HD95) and Mean Surface Distance (MSD), which capture worst-case and average deviations between predicted and reference contours, respectively (Ye et al., 2024). In addition, pixel-level metrics such as Root Mean Square Error (RMSE) assess decomposition precision (Wang et al., 2025).

Clinical (Retrieval, Classification, and Synthesis) Metrics. In image–text retrieval, Recall@K measures whether the correct reference appears among the top-K results (Xin et al., 2025). Classification performance is commonly assessed using Accuracy, F1-score, and AUC (Lu et al., 2025; Zhao et al., 2025). For text-guided image synthesis (e.g., generating tumor-bearing CT volumes from clinical descriptions), Root Mean Square Error (RMSE) and Pearson correlation are used to verify anatomical consistency between generated and reference images (Chen et al., 2025a). Additionally, in speech-oriented models, Medical Term Recognition Rate (MTRR) is used to ensure critical medical terminology is preserved (Lv et al., 2025).

Lexical/NLP Metrics. For report generation and VQA, standard automatic metrics such as BLEU, ROUGE, METEOR, and CIDEr are commonly used to quantify lexical similarity between generated text and reference reports. However, these surface-level metrics often fail to reflect the clinical correctness and diagnostic validity (Liu et al., 2024; Luo et al., 2025). To address these limitations, `clinically-oriented evaluation measures` have been increasingly adopted. Early work computes precision, recall, and F1 scores over extracted clinical findings using clinical labelers such as RadBERT (Yan et al., 2022). More recent metrics like GREEN (Ostmeier et al., 2024) and RaTEScore (Zhao et al., 2024) provide a finer-grained evaluation of radiological accuracy and report completeness (Liu et al., 2024).

Human Evaluation. Beyond automated metrics, human evaluation remains the gold standard for assessing generated clinical text, with experts rating generated reports for correctness, completeness, and conciseness (Batool et al., 2025). However, evaluation protocols vary: some employ Likert-scale scoring to quantify clinical accuracy (Liu et al., 2024; Batool et al., 2025), whereas others utilize pairwise ranking in which annotators select the best report among competing systems (Liu et al., 2021b). More recently, LLMs have been proposed as automated judges, but evidence suggests that they may systematically favor AI-generated text over human reference, highlighting the continued need for human evaluation (Batool et al., 2025).

9 Applications

3D-MedVLMs have been adapted for a diverse array of clinical downstream tasks, which we categorize into five primary domains (Table 1).

Diagnostic Classification and Anatomical Identification is a primary application of 3D-MedVLMs in medical imaging. These models extend beyond binary detection to recognize fine-grained recognition of disease subtypes and spatial localization (Gao et al., 2025; Lu et al., 2025). For example, RadCLIP (Lu et al., 2025), a contrastive-based model, demonstrates strong zero-shot generalization. To address data scarcity, Petersen et al. (2025) shows that scaling the number of negative pairs enables training with 62 scans. In contrast, task-specific architectures prioritize fine-grained optimization and yield higher performance on targeted tasks (Zhao et al., 2025; Wang et al., 2025).

Automated Report Generation converts visual findings to clinical narratives. Models such as 3DLVR (Du et al., 2024a) and Medical-VLBERT (Liu et al., 2021b) align visual and linguistic features across modalities. Recent work has extended report generation to dynamic or time-sensitive applications, including spine CT (Batool et al., 2025) and emergency head CT (Tomassini et al., 2025). A central challenge is hallucination, where the model generates clinically incorrect statements. Consequently, evaluation practices have shifted from surface-level lexical metrics to clinically grounded measures (Liu et al., 2024).

Semantic Segmentation and Visual Grounding seeks to align textual findings with anatomical coordinates in volumetric images, which can improve spatial localization and downstream task performance when incorporated in models (Luo et al., 2025). When traditional visual prompts (e.g., points or bounding boxes) are ineffective, particularly for anatomies with sparse or fragmented tissue, text-prompted pretraining has been proposed to unify heterogeneous datasets and guide segmen-

tation via semantic descriptions (Ye et al., 2024).

Image Synthesis generates anatomically accurate visual pathologies from textual clinical descriptions. By conditioning generative models on structured metadata, these approaches help mitigate data scarcity while maintaining adherence to medical guidelines. For example, diffusion-based methods have been used to synthesize renal tumors from nephrometry scores, allowing granular control over tumor morphology (Chen et al., 2025a). Similarly, adversarial frameworks have been applied to cross-modality tasks; MultiModalGAN synthesizes chest X-rays directly from text reports, while CycleGAN has been utilized to convert MRI scans into CT-like images, providing diagnostic alternatives when CT is unavailable (Ghadekar et al., 2025).

Retrieval Tasks have recently shifted from purely technical optimization to enabling clinically robust applications and overcoming data scarcity. Med3DVLM (Xin et al., 2025) demonstrated that using pairwise sigmoid loss during contrastive learning removed the need for large negative batches and outperformed the SOTA model by 19%. Complementing these model-centric advances, Yamamoto and Kikuchi (2025) tackled the bottleneck of dataset construction by incorporating lesion- and organ-aware supervision. Their method achieved a Top-1 accuracy of 51.7%, demonstrating the feasibility of slice-based retrieval for automating the curation of large-scale medical VQA datasets from routine clinical archives.

10 Challenges and Future Direction

While 3D medical VLMs have demonstrated impressive capabilities, they face challenges before they can be widely adopted as clinical assistants. We identify five key areas where future efforts need be directed: reliability, interpretability, data ecosystems, clinical alignment, and scalability.

The major barrier to deploying 3D-MedVLMs is ensuring `validity and reliability`. Current models remain prone to `hallucinations`, where models generate plausible but incorrect medical assertions (Qin et al., 2024). This issue is exacerbated in 3D-MedVLMs because radiologists often rely on external information (e.g., prior reports), and models trained on such data are more prone to hallucination (Luo et al., 2025). Moreover, maintaining precise feature alignment in complex anatomical scenarios is also challenging, as simple approaches often struggle with large 3D batches or rich cross-modal interactions (Xin et al., 2025). Furthermore, models often degrade under `domain shift`, with variations in scanner protocols or imaging artifacts significantly altering model behavior. Future work should focus on robustness to these shifts, ensuring models can generalize from curated datasets to the real clinical environments (Hamamci et al., 2025b).

Clinical deployment requires `transparency and explainability`, as trust relies on the ability to audit the decision-making process. Future architectures may emphasize `glass-box` designs to offer transparent, guideline-aligned reasoning (Gai et al., 2025). Promising directions include chain-of-radiology-thought prompting combined with visual grounding that links textual findings to voxel-level evidence (Luo et al., 2025). Equally important is integrating `uncertainty` to identify low-confidence predictions and reduce overconfident errors.

The third priority is building `healthy data ecosystems.` Although hospitals generate abundant raw data, high-quality, instruction-tuned 3D pairs remain scarce. Future efforts should focus on self-sustaining `automatic data pipelines` that reduce reliance on manual annotation, for example, using LLMs to transform noisy clinical reports into machine-readable formats at scale (Zhang et al., 2025). Simultaneously, data access and diversity constraints must be addressed, and `synthetic data generation` may be leveraged to augment rare pathologies, thereby improving model robustness and fairness for long-tail conditions (Ghadekar et al., 2025; Chen et al., 2025a; Lv et al., 2025).

Another direction is `clinical alignment`, as a gap exists between standard technical evaluation and clinical utility (Liu et al., 2024; Luo et al., 2025). Qin et al. (2024) shows that high scores on lexical metrics (e.g., BLEU, ROUGE) do not guarantee clinically accurate reports, obscuring true model readiness. To address this, the field should establish standardized, clinically rigorous benchmarks that reflect real-world workflows (e.g., diagnostic efficiency and clinician acceptance).

Finally, `scalability` remains a major bottleneck. The high computational cost of 3D data processing limits the deployment of 3D-MedVLMs in resource-constrained clinical environments (Xin et al., 2025). Strategies such as downsampling reduce computational overhead but inevitably sacrifice spatial details (Xin et al., 2025). Future work should explore parameter-efficient mechanisms to maintain diagnostic fidelity while lowering the computational overhead (Chen et al., 2025b).

Limitations

While this survey provides a broad and systematic overview of recent VLLM research, our analysis is necessarily qualitative rather than exhaustively quantitative. The rapid evolution and wide scope of the field make detailed head-to-head benchmarking difficult, particularly given the heterogeneity of evaluation metrics, datasets, and experimental settings across studies. As a result, performance comparisons are presented primarily in terms of high-level trends rather than rigorous quantitative meta-analysis. In addition, our review is limited to English-language publications and prioritizes peer-reviewed work from major conferences such as ACL, CVPR, and MICCAI. Also, influential arXiv preprints and some emerging or non-English contributions may have been overlooked. Finally, the fast pace of industrial and academic progress means that certain cutting-edge methods may not yet be fully captured. In future work, we plan to incorporate more quantitative meta-analysis where feasible, integrate insights from large-scale industrial deployments, and continuously update the survey to reflect ongoing advances in this rapidly developing field.

References

Michael J Ackerman. 2022. The visible human project. *Inf. Serv. Use*, 42(1):129–136.

Parnian Afshar, Shahin Heidarian, Nastaran Enshaei, Farnoosh Naderkhani, Moezedin Javad Rafiee, Anastasia Oikonomou, Faranak Babaki Fard, Kaveh Samimi, Konstantinos N Plataniotis, and Arash Mohammadi. 2021. COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning. *Sci. Data*, 8(1):121.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. 2020. Dataset of breast ultrasound images. *Data Brief*, 28(104863):104863.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: A visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing System*, pages 23716–23736.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F Christ, Richard K G Do, Marc J Gollub, Stephan H Heckers, Henkjan Huisman, William R Jarnagin, and 39 others. 2022. The medical segmentation decathlon. *Nat. Commun.*, 13(1):4128.

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

Humaira Batool, Asmat Mukhtar, Sajid Gul Khawaja, Norah Saleh Alghamdi, Asad Mansoor Khan, Adil Qayyum, Ruqqayia Adil, Zawar Khan, Muhammad Usman Akram, Muhammad Usman Akbar, and Anders Eklund. 2025. Knowledge distillation and transformer-based framework for automatic spine CT report generation. *IEEE Access*, 13:42949–42964.

Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Muller. 2024. VQA-med: Overview of the medical visual question answering task at ImageCLEF 2019.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohe, and 18 others. 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Trans. Med. Imaging*, 37(11):2514–2525.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, Fabian Lohöfer, Julian Walter Holch, Wieland Sommer, Felix Hofmann, Alexandre Hostettler, Naama Lev-Cohain, Michal Drozdzal, Michal Marianne Amitai, Refael Vivanti, and 90 others. 2023. The liver tumor segmentation benchmark (LiTS). *Med. Image Anal.*, 84(102680):102680.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13590–13618, Stroudsburg, PA, USA. Association for Computational Linguistics.

Victor M Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, Mario Parreno, Alberto Albiol, Fanwei Kong, Shawn C Shadden, Jorge Corral Acero, Vaanathi Sundaresan, Mina Saber, Mustafa Elattar, Hongwei Li, and 28 others. 2021. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. *IEEE Trans. Med. Imaging*, 40(12):3543–3554.

Tien-Yu Chang, Qinglin Gou, Leyi Zhao, Tiancheng Zhou, Hongyu Chen, Dong Yang, Huiwen Ju, Kaleb E Smith, Chengkun Sun, Jinqian Pan, Yu Huang, Xing He, Xuhong Zhang, Daguang Xu, Jie Xu, Jiang Bian, and Aokun Chen. 2025. From image to report: automating lung cancer screening interpretation and reporting with vision-language models. *J. Biomed. Inform.*, 171(104931):104931.

Huan-Yu Chen, Woan-Shiuan Chien, Ching-Heng Lin, and Chi-Chun Lee. 2025a. Leveraging foundation models for clinically instructed tumor image synthesis in renal cell carcinoma. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Junhao Chen, Zhiyuan Ding, Yan Liu, Xiangzhu Zeng, and Ling Wang. 2025b. FLIQA-AD: A fusion model with large language model for better diagnose and MMSE prediction of alzheimer's disease. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 587–594, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qiuhui Chen, Xuancheng Yao, Huping Ye, and Yi Hong. 2025c. Enhancing 3D medical image understanding with pretraining aided by 2D multimodal large language models. *arXiv [cs.CV]*, pages 1–14.

Sihong Chen, Kai Ma, and Yefeng Zheng. 2019. Med3D: Transfer learning for 3D medical image analysis. *arXiv [cs.CV]*.

Xiaojun Chen, Jia Ke, Yaning Zhang, Jianping Gou, Anna Shen, and Shaohua Wan. 2025d. Multimodal distillation pre-training model for ultrasound dynamic images annotation. *IEEE J. Biomed. Health Inform.*, 29(5):3124–3136.

Yinda Chen, Che Liu, Xiaoyu Liu, Rossella Arcucci, and Zhiwei Xiong. 2024. BIMCV-R: A landmark dataset for 3D CT text-image retrieval. In *Lecture Notes in Computer Science*, Lecture Notes in Computer Science, pages 124–134. Springer Nature Switzerland, Cham.

Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. 2018. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet*, 392(10162):2388–2396.

Stephanie J Chiu, Michael J Allingham, Priyatham S Mettu, Scott W Cousins, Joseph A Izatt, and Sina Farsiu. 2015. Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. *Biomed. Opt. Express*, 6(4):1172–1194.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *arXiv [cs.LG]*.

Ethan Dack, Lorenzo Brigato, Matthew McMurray, Matthias Fontanellaz, Thomas Frauenfelder, Hanno Hoppe, Aristomenis Exadaktylos, Thomas Geiser, Manuela Funke-Chambour, Andreas Christe, Lukas Ebner, and Stavroula Mougiakakou. 2023. An empirical analysis for zero-shot multi-label classification on COVID-19 CT scans and uncurated reports. *arXiv [eess.IV]*, pages 2626–2635.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE.

Hanyue Du, Yong Zhuang, and Jingyu Xu. 2024a. Symbiotic fusion of evolving anatomical landscapes for intelligent diagnostic narratives. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4513–4520. IEEE.

Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. 2024b. Segvol: Universal and interactive volumetric medical image segmentation. *Adv. Neural Inf. Process. Syst.*, 37:110746–110783.

Sina Farsiu, Stephanie J Chiu, Rachelle V O'Connell, Francisco A Folgar, Eric Yuan, Joseph A Izatt, Cynthia A Toth, and Age-Related Eye Disease Study 2 Ancillary Spectral Domain Optical Coherence Tomography Study Group. 2014. Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology*, 121(1):162–172.

Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. MedThink: A rationale-guided framework for explaining medical visual question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7438–7450, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yaozong Gao, Yiran Shu, Mingyang Yu, Yanbo Chen, Jingyu Liu, Shaonan Zhong, Weifang Zhang, Yiqiang Zhan, Xiang Sean Zhou, Xinlu Wang, Meixin Zhao, and Dinggang Shen. 2025. Hierarchical contrastive learning for precise whole-body anatomical localization in PET/CT imaging. *IEEE Trans. Med. Imaging*, PP:1–1.

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

Premanand Ghadekar, Pranav Jadhav, Sai Kulkarni, Isha Kulkarni, Om Lohade, and Isha Mahajan. 2025. Multimodal medical image synthesis and cross-modality conversion for improved diagnosis and explainable AI analysis. In *2025 International Conference on Artificial intelligence and Emerging Technologies (ICAIET)*, pages 1–6. IEEE.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv [cs.AI]*.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, and 11 others. 2025a. Developing generalist foundation models from a multimodal dataset for 3D computed tomography. *arXiv [cs.CV]*.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, and 11 others. 2025b. Developing generalist foundation models from a multimodal dataset for 3D computed tomography. *arXiv [cs.CV]*.

Seyed Mohammad Hossein Hashemi. 2023. Crystal clean: Brain tumors MRI dataset.

Taimur Hassan, M Usman Akram, M Furqan Masood, and Ubaidullah Yasin. 2018. BIOMISA retinal image database for macular and ocular syndromes. In *Lecture Notes in Computer Science*, Lecture Notes in Computer Science, pages 695–705. Springer International Publishing, Cham.

Taimur Hassan, Hina Raja, Kais Belwafi, Samet Akcay, Mohamed Jleli, Bessem Samet, Naoufel Werghi, Jawad Yousaf, and Mohammed Ghazal. 2025. A vision language correlation framework for screening disabled retina. *IEEE J. Biomed. Health Inform.*, 29(2):1283–1296.

Xuehai He, Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. 2020. MedDialog: Two large-scale medical dialogue datasets. *arXiv [cs.LG]*.

Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, Yoel Shoshan, Flora Gilboa-Solomon, Yasmeen George, Xi Yang, Jianpeng Zhang, Jing Zhang, Yong Xia, Mengran Wu, Zhiyang Liu, and 26 others. 2023. The KiTS21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT. *arXiv [cs.CV]*.

Shishuai Hu, Zehui Liao, and Yong Xia. 2022. Domain specific convolution and high frequency reconstruction based unsupervised domain adaptation for medical image segmentation. In *Lecture Notes in Computer Science*, Lecture Notes in Computer Science, pages 650–659. Springer Nature Switzerland, Cham.

Shih-Cheng Huang, Zepeng Huo, E Steinberg, Chia-Chun Chiang, Curtis P Langlotz, M Lungren, Serena Yeung, Nigam H Shah, and Jason Alan Fries. 2023. INSPECT: A multimodal dataset for patient outcome prediction of pulmonary embolisms. *Neural Inf Process Syst*, 36:17742–17772.

Akimichi Ichinose, Taro Hatsutani, Keigo Nakamura, Yoshiro Kitamura, Satoshi Iizuka, Edgar Simo-Serra, Shoji Kido, and Noriyuki Tomiyama. 2023. Visual grounding of whole radiology reports for 3D CT images. *arXiv [cs.CV]*, pages 611–621.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A Mong, Safwan S Halabi, Jesse K Sandberg, Ricky Jones, David B Larson, Curtis P Langlotz, Bhavik N Patel, Matthew P Lungren, and Andrew Y Ng. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv [cs.CV]*, pages 590–597.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv [cs.CV]*, pages 4904–4916.

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

Mansi Kakkar, Dattesh Shanbhag, Chandan Aladahalli, and Gurunath Reddy. 2024. Language augmentation in CLIP for improved anatomy detection on multimodal medical images. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4. IEEE.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137. IEEE.

Pratham Kaushik, Vinay Kukreja, Eshika Jain, Modafar Ati, and Shanmugasundaram Hariharan. 2025. Neuroinfinity: A lightweight prompt-driven vision–language framework for alzheimer's disease stage classification using magnetic resonance imaging. *IEEE Access*, 13(99):191589–191608.

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, Debdoot Sheet, and 8 others. 2021. CHAOS challenge - combined (CT-MR) healthy abdominal organ segmentation. *Med. Image Anal.*, 69(101950):101950.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina C S Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K Prasadha, Jacqueline Pei, Magdalene Y L Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, and 20 others. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9.

Soo Yong Kim and Seunghyeok Hong. 2026. Enhancing 3D medical vision-language models with slice-wise visual prompt. In *Lecture Notes in Computer Science*, Lecture Notes in Computer Science, pages 197–205. Springer Nature Switzerland, Cham.

Lauren N Koenig, Gregory S Day, Amber Salter, Sarah Keefe, Laura M Marple, Justin Long, Pamela LaMontagne, Parinaz Massoumzadeh, B Joy Snider, Manasa Kanthamneni, and Others. 2020. Select atrophied regions in alzheimer disease (SARA): An improved volumetric model for identifying alzheimer disease dementia. *NeuroImage Clin.*, 26:102248.

Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. 2024. Language guided domain generalized medical image segmentation. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Alain Lalande, Zhihao Chen, Thomas Decourselle, Abdul Qayyum, Thibaut Pommier, Luc Lorgis, Ezequiel de la Rosa, Alexandre Cochet, Yves Cottin, Dominique Ginhac, Michel Salomon, Raphaël Couturier, and Fabrice Meriaudeau. 2020. Emidec: A database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac MRI. *Data (Basel)*, 5(4):89.

Pamela J LaMontagne, Tammie L S Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, and Others. 2019. OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medRxiv*, pages 2019–2012.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Sci. Data*, 5(1):180251.

Cheng-Yi Li, Kao-Jung Chang, Cheng-Fu Yang, Hsin-Yu Wu, Wenting Chen, Hritik Bansal, Ling Chen, Yi-Ping Yang, Yu-Chun Chen, Shih-Pin Chen, Shih-Jen Chen, Jiing-Feng Lirng, Kai-Wei Chang, and Shih-Hwa Chiou. 2025a. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. *Nat. Commun.*, 16(1):2258.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv [cs.CV]*, pages 19730–19742.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv [cs.CV]*, pages 12888–12900.

Minghan Li, Congcong Wen, Yu Tian, Min Shi, Yan Luo, Hao Huang, Yi Fang, and Mengyu Wang. 2025b. FairFedMed: Benchmarking group fairness in federated medical imaging with FairLoRA. *IEEE Trans. Med. Imaging*, PP:1–1.

Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro R A S Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, Yutong Tang, Yining Cao, Haoqi Han, Zheyuan Zhang, Jiawei Liu, Tiezheng Zhang, Yujiu Ma, Jincheng Wang, Guang Zhang, and 2 others. 2025c. AbdomenAtlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *arXiv [cs.CV]*, page 103285.

Yingwei Li, Zhuotun Zhu, Yuyin Zhou, Yingda Xia, Wei Shen, Elliot K Fishman, and Alan L Yuille. 2019. Volumetric medical image segmentation: A 3D deep coarse-to-fine framework and its adversarial examples. In *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*, Advances in Computer Vision and Pattern Recognition, pages 69–91. Springer International Publishing, Cham.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021a. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Che Liu, Zhongwei Wan, Yuqi Wang, Hui Shen, Haozhe Wang, Kangyu Zheng, Mi Zhang, and Rossella Arcucci. 2024. Argus: Benchmarking and enhancing vision-language models for 3D radiology report generation. *arXiv [cs.CV]*.

Guangyi Liu, Yinghong Liao, Fuyu Wang, Bin Zhang, Lu Zhang, Xiaodan Liang, Xiang Wan, Shaolin Li, Zhen Li, Shuixing Zhang, and Shuguang Cui. 2021b. Medical-VLBERT: Medical visual language BERT

for COVID-19 CT report generation with alternate learning. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(9):3786–3797.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv [cs.CV]*, pages 34892–34916.

Jingxin Liu, Xinran Zhu, Zhangzhen Shi, Donghong An, Lihui Zu, Kailiang Cheng, and Xiaopeng Guo. 2025. Text-guided multimodal deep learning in magnetic resonance imaging for spinal structures segmentation and lumbar abnormalities identification. *Quant. Imaging Med. Surg.*, 15(10):9710–9728.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2021c. Video swin transformer. *arXiv [cs.CV]*, pages 3202–3211.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv [cs.CV]*, pages 13–23.

Zhixiu Lu, Hailong Li, Nehal A Parikh, Jonathan R Dillman, and Lili He. 2025. RadCLIP: Enhancing radiologic image analysis through contrastive language-image pretraining. *IEEE Trans. Neural Netw. Learn. Syst.*, 36(10):17613–17622.

Lingxiao Luo, Bingda Tang, Xuanzhong Chen, Rong Han, and Ting Chen. 2025. VividMed: Vision language model with versatile visual grounding for medicine. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1800–1821, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N Metaxas, Guotai Wang, and Shaoting Zhang. 2022. WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. *Med. Image Anal.*, 82(102642):102642.

Jianhui Lv, Wadii Boulila, Shalli Rani, and Huamao Jiang. 2025. Enhanced multimodal speech processing for healthcare applications: A deep fusion approach. *IEEE J. Sel. Top. Signal Process.*, 19(4):600–612.

Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. 2022. AbdomenCT-1K: Is abdominal organ segmentation a solved problem? *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):6695–6714.

Manal Makram and Ammar Mohammed. 2025. Enhancing large vision language models for liver CT scans in medical reports. In *2025 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 88–95. IEEE.

Ian B Malone, David Cash, Gerard R Ridgway, David G MacManus, Sebastien Ourselin, Nick C Fox, and Jonathan M Schott. 2013. MIRIAD–public release of a multiple time point alzheimer's MR imaging dataset. *Neuroimage*, 70:33–36.

Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. 2010. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *J. Cogn. Neurosci.*, 22(12):2677–2684.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. 2007. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.*, 19(9):1498–1507.

Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, and Others. 2011. The parkinson progression marker initiative (PPMI). *Prog. Neurobiol.*, 95(4):629–635.

Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A Fayad, and Yang Yang. 2022. RadImageNet: An open radiologic deep learning research dataset for effective transfer learning. *Radiol. Artif. Intell.*, 4(5):e210315.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 79–87. IEEE.

Asbjørn Munk, Jakob Ambsdorf, Sebastian Llambias, and Mads Nielsen. 2024. AMAES: Augmented masked autoencoder pretraining on public brain MRI data for 3D-native segmentation. *arXiv [eess.IV]*.

Maiko Nagao, Atsushi Teramoto, Kaito Urata, Kazuyoshi Imaizumi, Masashi Kondo, and Hiroshi Fujita. 2025. Preliminary study on image-finding generation and classification of lung nodules in chest CT images using vision–language models. *Computers*, 14(11):489.

Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Q Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. 2024. Unveiling multilevel and multi-modal semantic representations in the human brain using large language models. *bioRxiv*, pages 20313–20338.

Friska Natalia, Hira Meidia, Nunik Afriliana, Ala S Al-Kafri, Sud Sudirman, Andrew Simpson, Ali Sophian, Mohammed Al-Jumaily, Wasfi Al-Rashdan, and Mohammad Bashtawi. 2018. Development of ground truth data for automatic lumbar spine MRI image

segmentation. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1449–1454. IEEE.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang T T Tong, Diep H Dinh, Cuong D Do, Luu T Doan, Cuong N Nguyen, Binh T Nguyen, Que V Nguyen, Au D Hoang, Hien N Phan, Anh T Nguyen, Phuong H Ho, and 5 others. 2022. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Sci. Data*, 9(1):429.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2023. GPT-4 technical report. *arXiv [cs.CL]*, page 08774.

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

Shumao Pang, Chunlan Pang, Lei Zhao, Yangfan Chen, Zhihai Su, Yujia Zhou, Meiyan Huang, Wei Yang, Hai Lu, and Qianjin Feng. 2021. SpineParseNet: Spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation. *IEEE Trans. Med. Imaging*, 40(1):262–273.

Jakob Krogh Petersen, Valdemar Licht, Mads Nielsen, and Asbjørn Munk. 2025. Revisiting clip: Efficient alignment of 3D MRI and tabular data using domain-specific foundation models. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, C R Jack, William J Jagust, Leslie M Shaw, Arthur W Toga, and Others. 2010. Alzheimer's disease neuroimaging initiative (ADNI) clinical characterization. *Neurology*, 74(3):201–209.

Zhenyue Qin, Yu Yin, Dylan Campbell, Xuansheng Wu, Ke Zou, Yih-Chung Tham, Ninghao Liu, Xiuzhen Zhang, and Qingyu Chen. 2024. LMOD: A large multimodal ophthalmology dataset and benchmark for large vision-language models. *arXiv [cs.CV]*, pages 2501–2522.

Junyi Qiu, Lei Li, Sihan Wang, Ke Zhang, Yinyin Chen, Shan Yang, and Xiahai Zhuang. 2022. MyoPS-net: Myocardial pathology segmentation with flexible combination of multi-sequence CMR images. *arXiv [eess.IV]*, page 102694.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2024. Qwen2.5 technical report. *arXiv [cs.CL]*.

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. 2021. Learning transferable visual models from natural language supervision. *arXiv [cs.CV]*, pages 8748–8763.

Reza Rasti, Hossein Rabbani, Alireza Mehridehnavi, and Fedra Hajizadeh. 2018. Macular OCT classification using a multi-scale convolutional neural network ensemble. *IEEE Trans. Med. Imaging*, 37(4):1024–1034.

Tyler Richards, Jason Talbott, Robyn Ball, Errol Colak, Adam Flanders, Felipe Kitamura, John Mongan, Luciano Prevedello, and Maryam Vazirabad. 2024. RSNA 2024 lumbar spine degenerative classification. `https://kaggle.com/competitions/rsna-2024-lumbar-spine-degenerative-classification`.

Haruki Sakajo, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. Tonguescape: Exploring language models understanding of vowel articulation. *arXiv [cs.CL]*.

Anjany Sekuboyina, Malek E Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, Martin Urschler, Maodong Chen, Dalong Cheng, Nikolas Lessmann, Yujin Hu, Tianfu Wang, Dong Yang, Daguang Xu, Felix Ambellan, and 50 others. 2021. VerSe: A vertebrae labelling and segmentation benchmark for multi-detector CT images. *Med. Image Anal.*, 73(102166):102166.

Yanzhao Shi, Junzhong Ji, Xiaodan Zhang, Liangqiong Qu, and Ying Liu. 2023. Granularity matters: Pathological graph-driven cross-modal alignment for brain CT report generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6630, Stroudsburg, PA, USA. Association for Computational Linguistics.

Emil Y Sidky and Xiaochuan Pan. 2024. Report on the AAPM deep-learning spectral CT grand challenge. *Med. Phys.*, 51(2):772–785.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. 2025. Can VLMs actually see and read? a survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages

24452–24470, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pratul P Srinivasan, Leo A Kim, Priyatham S Mettu, Scott W Cousins, Grant M Comer, Joseph A Izatt, and Sina Farsiu. 2014. Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. *Biomed. Opt. Express*, 5(10):3568–3577.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tianchi. 2020. Spinal disease dataset. https://tianchi.aliyun.com/dataset/dataDetail?dataId=79463.

Selene Tomassini, Damiano Duranti, Abdallah Zeggada, Carlo Cosimo Quattrocchi, Farid Melgani, and Paolo Giorgini. 2025. Multi-branch CNN-LSTM fusion network-driven system with BERT semantic evaluator for radiology reporting in emergency head CTs. *IEEE J. Transl. Eng. Health Med.*, 13:61–74.

Andrea C Tricco, Erin Lillie, Wasifa Zarin, Kelly K O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah D J Peters, Tanya Horsley, Laura Weeks, Susanne Hempel, Elie A Akl, Christine Chang, Jessie McGowan, Lesley Stewart, Lisa Hartling, Adrian Aldcroft, Michael G Wilson, Chantelle Garritty, and 9 others. 2018. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Ann. Intern. Med.*, 169(7):467–473.

Santiago Vitale, José Ignacio Orlando, Emmanuel Iarussi, and Ignacio Larrabide. 2020. Improving realism in patient-specific abdominal ultrasound simulation using CycleGANs. *Int. J. Comput. Assist. Radiol. Surg.*, 15(2):183–192.

Xianghong Wang, Jiajun Xiang, Aihua Mao, Jiayi Xie, Peng Jin, Mingchao Ding, Yixuan Yuan, Yanye Lu, Lequan Yu, Hongmin Cai, Baiying Lei, and Tianye Niu. 2025. Clip-driven universal model for multi-material decomposition in dual-energy CT. *IEEE Trans. Comput. Imaging*, 11:349–361.

Yibin Wang, William Duggar, David Caballero, Toms Vengaloor Thomas, Neha Adari, Eswarakumar Mundra, and Haifeng Wang. 2023. Brain tumor recurrence prediction after gamma knife radiotherapy from MRI and related DICOM-RT: An open annotated dataset and baseline algorithm (brain-TR-GammaKnife).

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. MedCLIP: Contrastive learning from unpaired medical images and text. *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2022:3876–3887.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. 2023. TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiol. Artif. Intell.*, 5(5):e230024.

Navodini Wijethilake, Marina Ivory, Oscar MacCormac, Siddhant Kumar, Aaron Kujawa, Lorena Garcia-Foncillas Macias, Rebecca Burger, Amanda Hitchings, Suki Thomson, Sinan Barazi, Eleni Maratos, Rupert Obholzer, Dan Jian, Fiona McClenaghan, Kazumi Chia, Omar Al-Salihi, Nick Thomas, Steve Connor, Tom Vercauteren, and Jonathan Shapey. 2025. Deep learning consensus-based annotation of vestibular schwannoma from magnetic resonance imaging: An annotated multi-center routine clinical dataset (vestibular-schwannoma-MC-RC 2).

Fuping Wu and Xiahai Zhuang. 2020. CF distance: A new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation. *IEEE Trans. Med. Imaging*, 39(12):4274–4285.

Jing Wu, Yuli Wang, Zhusi Zhong, Weihua Liao, Natalia Trayanova, Zhicheng Jiao, and Harrison X Bai. 2025. Vision-language foundation model for 3D medical imaging. *NPJ Artif. Intell.*, 1(1):15.

Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. 2025. Med3DVLM: An efficient vision-language model for 3D medical image analysis. *arXiv [cs.CV]*, pages 1–14.

Kohei Yamamoto and Tomohiro Kikuchi. 2025. Feasibility study of CLIP-based key slice selection in CT images and performance enhancement via lesion- and organ-aware fine-tuning. *Bioengineering (Basel)*, 12(10):1093.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. 2022. RadBERT: Adapting transformer-based language models to radiology. *Radiol. Artif. Intell.*, 4(4):e210258.

Jiancheng Yang, Xiaoyang Huang, Yi He, Jingwei Xu, Canqian Yang, Guozheng Xu, and Bingbing Ni. 2021. Reinventing 2D convolutions for 3D images. *IEEE J. Biomed. Health Inform.*, 25(8):3009–3018.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. MedMNIST v2 - a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Sci. Data*, 10(1):41.

Huping Ye, Yushan Deng, and Yi Hong. 2024. Epicardial adipose tissue segmentation in MRIs using text-prompted pretraining model. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 7021–7028. IEEE.

Joonhyeok Yoon, Hangyeol Park, Minjun Kim, Hwihun Jeong, Se Young Chun, Sooyeon Ji, and Jongho Lee. 2025. Clinical dementia rating classification using integrated vision and language information. *IEEE Access*, 13:184602–184617.

Jianpeng Zhang, Yutong Xie, Yan Wang, and Yong Xia. 2021. Inter-slice context residual learning for 3D medical image segmentation. *IEEE Trans. Med. Imaging*, 40(2):661–672.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Weiwei Tian, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2025. Development of a large-scale grounded vision language dataset for chest CT analysis. *Sci. Data*, 12(1):1636.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv [cs.CV]*.

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

Yufang Zhao, Yue Li, Yanjing Zhang, Xiaohui Yan, Guolin Yin, and Liping Liu. 2025. Enhancing thyroid nodule assessment with UTV-ST swin kansformer: A multimodal approach to predict invasiveness. *IEEE Access*, 13:29081–29090.

Xiahai Zhuang and Juan Shen. 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. *Med. Image Anal.*, 31:77–87.

A Search Query Variables

```
{3d_terms} = "CT"
    OR "MRI"
    OR "3D"
    OR "volumetric"
    OR "volume"
    OR "computed tomography"
    OR "magnetic resonance"
    OR "3D image"
    OR "biomedical image"
    OR "3D reconstruction"
    OR "point cloud"

{model_terms} = "multimodal"
    OR "vision-language"
    OR "foundation model"
    OR "self-supervised"
    OR "contrastive"
    OR "CLIP"
    OR "LVLM"
    OR "image-text"
    OR "transformer"
    OR "pretrain*"
    OR "masked autoencoder"
    OR "cross-modal"
    OR "report-grounded"
    OR "retrieval"

{text_terms} = "radiology report"
    OR "clinical text"
    OR "report generation"
    OR "text"
    OR "language"
    OR "PACS"
    OR "DICOM"
    OR "diagnosis"
    OR "medical report"
    OR "biomedical text"

{video_terms} = "medical video"
    OR "surgical video"
    OR "ultrasound video"
    OR "endoscopy video"
    OR "cine MRI"
    OR "temporal CT"
```

Table 1: Overview of Included Studies.

Study	Code	Visionmodel	LLM	3D handling	Inst follow	Pretrained	Fine-tuned	Input	Task
VividMed (2025)	Link	ViT, SAM	Vicuna	Voxel	✓	E2E	E2E	CT, X-ray	Q R S
Argus (2024)	Link	ViT	Llama3	Voxel	✓	Vision	E2E	CT	R
Liu et al. (2025)	✗	ConvNeXt	MedCLIP	Slice	✗	Vision	E2E	MRI	S C
Zhang et al. (2025)	Link	3D ConvNet	Llama3	Voxel	✓	–	Language	CT	R Q
Yamamoto and Kikuchi (2025)	✗	ViT	CLIP	Slice	✗	–	E2E	CT	R
Tomassini et al. (2025)	Link	VGG16	LSTM, BERT	Slice	✗	–	Language	CT	R
Batool et al. (2025)	✗	ViT	BioBERT, MiniLM	Slice	✗	E2E	E2E	CT	R
FLIQA-AD (2025b)	Link	3D ViT	bioClinicalBert, FLAN-T5	Voxel	✓	–	E2E	MRI	C Q
Med3DVLM (2025)	Link	DCFromer	Qwen2.5-7B-Instruct	Voxel	✓	E2E	E2E	CT	Q C R R
RadCLIP (2025)	Link	ViT	CLIP	Slice	✗	E2E	Vision	CT, MRI	C R
Petersen et al. (2025)	✗	Swin-T, MedNeXt, ResNet	BERT	Voxel	✗	Vision	E2E	MRI	C R
MultiModalGAN (2025)	✗	CNN	BERT	Slice	✗	–	E2E	MRI, X-Ray	S
Med3DInsight (2025c)	✗	nnFormer	CLIP	Voxel	✗	E2E	E2E	CT, MRI	S C
Medical-VLBERT (2021b)	✗	DenseNet	BERT	Slice	✗	E2E	E2E	CT	R C
Chen et al. (2025a)	✗	ViT-H SAM	LLaVA-Med	Slice	✓	E2E	E2E	CT	S S
Ye et al. (2024)	✗	ViT	BERT	Slice	✗	Vision	Vision	MRI	S
Makram and Mohammed (2025)	✗	ViT, CLIP	Vicuna, Mistral, MPT, LLaMA, GPT-4o	Slice	✓	–	–	CT, MRI	C R Q
PMC-MSA (2024)	✗	ViT	CLIP	Slice	✗	–	E2E	CT, MRI	C
Dack et al. (2023)	✗	ViT, Swin-T, ResNet	GPT2, BioClinicalBERT, PubMedBERT, RadBERT	Slice	✗	–	E2E	CT	C
HCL-AL (2025)	Link	3D ResNet	CLIP	Voxel	✗	Language	–	PET, CT	R
Wang et al. (2025)	✗	Siamese	CLIP	Slice	✗	Language	–	CT	S
TGCFA (2024)	✗	U-Net	CLIP	Slice	✗	–	Vision	CT, Fundus, MRI	S
3DLVR (2024a)	✗	Transformer	Transformer	Voxel	✗	–	E2E	CT	R
DILBERT (2025d)	✗	3D CNN	BERT	Video	✗	E2E	E2E	Ultrasound	R
Zhao et al. (2025)	✗	Video Transformer	Kansformer	Video	✗	–	E2E	Ultrasound	C
HAV-DF (2025)	✗	3D-Conv, ResNet	LSTM	Video	✗	–	E2E	Audio, Video	C
Tonguescape (2025)	✗	Gemini 1.5 Pro, GPT-4o, LLaVANeXT-Interleave, Phi-3.5-vision-instruct, Qwen2-VL-Instruct, VideoLLaMA2		Video	✓	–	E2E	MRI video	Q
Drama2brain (2024)	Link	DeiT, ResNet, GIT, BridgeTower, LLaVa-v1.5	Word2Vec, BERT, GPT2, OPT, Llama2	Video	Both	–	–	fMRI	C
Shi et al. (2023)	✗	ResNet101	–	Slice	✗	–	E2E	CT	R
LMOD (2024)	✗	LLaVA, InternVL, Yi-6B, Qwen	VILA, GPT-4o,	Slice	✓	–	E2E	CFP, LP, OCT, SS, SLO	C Q
LUMEN (2025)	✗	YOLOv9, LlaMA-3-VILA1.5, LlaMA3-VILA-M3	LlaMA-3-VILA1.5, LlaMA3-VILA-M3, LlaMA 3.3	Slice	✓	–	E2E	CT	C S R
Yoon et al. (2025)	✗	ViT	Transformer	Voxel	✗	E2E	E2E	MRI	C
BrainGPT (2025a)	Link	ViT	LLaMA	Slice	✓	–	E2E	CT	R
NeuroInfinity (2025)	✗	ViT	CLIP	Slice	✗	–	E2E	MRI	C

Continued on next page

Table 1: Overview of Included Studies. (Continued)

Study	Code	Visionmodel	LLM	3D handling	Inst follow	Pretrained	Fine-tuned	Input	Task
Hassan et al. (2025)	✗	ResNet-50, PVT, ViT	BERT, Word2Vec, XLNet	Slice	✗	–	E2E	OCT	C
Nagao et al. (2025)	✗	BLIP, GiT	BLIP, GiT	Slice	✗	–	E2E	CT	R C
VISTA-Prompt (2026)	✗	DCFormer	Qwen-2.5	Slice	✓	–	Language	CT	Q R

Note: R - Report Generation, Q - Visual Question Answering, C - Diagnostic Classification and Anatomical Identification, S - Segmentation and Visual Grounding, S - Image Synthesis and Modality Conversion, R - Retrieval, Inst Follow - Instruction following, E2E - End-to-end

Table 2: Overview of recent benchmark datasets.

Dataset Name	Data Modality	Annot. Type	Scale (# samples)	Public URL	Institution	Clinical Specialty
TotalSegmentator (2023)	CT	S	1,204	Link	Single	Clinical
VinDr-CXR (2022)	X-Ray	S C	18,000	Link	Multiple	Clinical
VQA-RAD (2018)	CT, MRI, X-Ray	Q	315	Link	Single	Clinical
SLAKE (2021a)	CT, MRI, X-Ray	Q	642	Link	Multiple	Clinical
VQA-Med (2024)	Angiography, CT, Mammograph, MRI, PET, X-Ray, Ultrasound	Q	4,200	Link	—	Clinical
MIMIC-CXR (2019)	X-Ray	C R	377,110	Link	Single	Clinical
CT-RATE (2025a)	CT	C R	25,692	Link	Single	Clinical
BIMCV-R (2024)	CT	R	8,069	Link	Single	Clinical
INSPECT (2023)	CT	C R	23,248	Link	Single	Pulmonary
MRSpineSeg2021 (2021)	MRI	S	215	Link	Single	Orthopedics
RSNA Lumbar Spine Degenerative (2024)	MRI	C	1,975	Link	Single	Orthopedics
Tianchi Spinal Disease (2020)	MRI	S C	201	Link	Single	Orthopedics
Lumbar Spine MRI dataset (2018)	MRI	S R	515	Link	Single	Orthopedics
RadGenome-Chest CT (2025)	CT (Chest)	S R Q	25,692	Link	Single	Clinical
VerSe20 (2021)	CT	—	300	Link	Single	Orthopedics
ADNI (2010)	MRI	C	8,315	Link	Multiple	Neurology
OASIS-2 (2010)	MRI	C	373	Link	Single	Neurology
rtMRIDB	MRI (Video)	C	26K	Link	Single	Otolaryngology
VowelVideo (2025)	MRI (Video)	C	120	Link	Single	Otolaryngology
drama2brain (2024)	fMRI	C	6 subjects	Link	Single	Neurology
LMOD (2024)	CFP, LP, OCT, SLO, SS	S C Q	21,993	Link	Multiple	Ophthalmology
OASIS-3 (2019)	MRI	C	2,842	Link	Single	Neurology
OASIS-4 (2020)	MRI	C	676	Link	Single	Neurology
CQ500 (2018)	CT	R	1,154	Link	Single	Neuroradiology
ImagesOASIS (2007)	MRI	C	80,00	Link	Single	Neurology
MIRIAD (2013)	MRI	C	708	Link	Single	Neurology
Zhang dataset (2018)	OCT	C	109,309	Link	Multiple	Ophthalmology
Duke-1 (2014)	OCT	C	38,300	Link	Multiple	Ophthalmology
Duke-2 (2015)	OCT	C	610	Link	Single	Ophthalmology
Duke-3 (2014)	OCT	C	3,231	Link	Multiple	Ophthalmology
Rabbani (2018)	OCT	C	4,241	Link	Single	Ophthalmology
BIOMISA (2018)	OCT	C	4,163	Link	Single	Ophthalmology
PMC-VQA (2023)	—	Q	149K	Link	Multiple	Clinical
M3D-Cap (2024b)	CT	Q	120k	Link	Multiple	Clinical
M3D-VQA (2024)	CT	Q	660K	Link	Multiple	Clinical
RadCLIP (2025)	CT, MRI, X-Ray	Q	1,210,083	Link	Multiple	Clinical
ChestXpert (2019)	X-Ray	C	224,316	Link	Single	Clinical
Crystal Clean Brain Tumor (2023)	MRI	C	3,264	Link	—	Neurosurgery
IXI Brain	MRI	C	600	Link	Multiple	Neurology
COVID-CT-MD (2021)	CT	C	23,349	Link	Single	Pulmonary

Continued on next page

Table 2: Overview of recent benchmark datasets. (Continued)

Dataset Name	Data Modality	Annot. Type	Scale (# samples)	Public URL	Institution	Clinical Specialty
BRAINS-45K (2024)	MRI	—	44,756	Link	Multiple	Neurology
Brain-TR-GammaKnife (2023)	MRI	Q	17,191	Link	Single	Neurosurgery
3DSeg-8 (2019)	CT, MRI	Q	25K	Link	Multiple	Clinical
MM-WHS (2016)	CT	S	20	Link	Single	Cardiology
MSD-Heart (2022)	CT, MRI	S	30	Link	—	Cardiology
CHAOS (2021)	CT, MRI	S	80	Link	Single	Clinical
AbdomenCT-1K (2022)	CT	S	1,112	Link	Multiple	Clinical
VS (2025)	MRI	S	676	Link	Multiple	Neurosurgery
LiTs (Bilic et al., 2023)	CT	S	201	Link	Single	Hepatology
PPMI (2011)	MRI	C	582	Link	Multiple	Neurology
COVID-19 CT (2021b)	CT	R	1,104	Link	Multiple	Pulmonary
AbdomenAtlas (2025c)	CT	S	3.2m	Link	Multiple	Clinical
WORD (2022)	CT	S	30,495	Link	Single	Clinical
KiTS23 (2023)	CT	S	599	Link	Single	Urology
ACDC (2018)	MRI	S	150	Link	Single	Cardiology
M&Ms (2021)	MRI	S	375	Link	Multiple	Cardiology
Wu and Zhuang (2020)	MRI	S	1,568	Link	Multiple	Cardiology
Emidec (2020)	MRI	S	150	Link	Single	Cardiology
MyoPS2020 (2022)	MRI	S	45	Link	Single	Cardiology
Eurorad	CT, MRI	R	5K	Link	—	Clinical
Visible Human Project (2022)	CT	C	650	Link	—	Research
Auto-PET (2022)	CT, PET	C	1,169	Link	Multiple	Clinical
DL-spectral CT (2024)	CT	S	1,100	Link	Synthetic	Clinical
MedMNIST (2023)	Dermatology, CT, MRI, OCT, Pathology, Retina, X-Ray, Ultrasound	C	718K	Link	Multiple	Clinical
MS-CMR	MRI	S	<1K	Link	—	Clinical
RIGA+ (2022)	Fundus	S	750	Link	Multiple	Ophthalmology
BUSI (2020)	Ultrasound	R	780	Link	Single	Breast Surgery
USSS (2020)	Ultrasound	R	926	Link	Synthetic	Clinical, Simulation
MedDialog (2020)	Audio	D	260K	Link	Multiple	Clinical
MedVidQA	Video	Q	6.4K	Link	Multiple	Clinical, Educational

Note: S - Segmentation Mask and Bounding Box, Q - Question-Answer Pair, C - Classification Label, R - Report, D - Dialogue