# Volumetric Radiology AI in the Era of Multimodal Large Language Models

**Zanting Ye[1*], Shengyuan Liu[2*†], Xin Liu[1], Chenhui Wang[3], Zhisong Wang[4], Jiashuai Liu[5], Zipei Wang[6], Cheng Wang[2], Wentao Pan[2], Mengjie Fang[6], Di Dong[6], Mohammad Salmanpour[7], Arman Rahmim[7], Yu Gu[8], Yong Xia[4], Hongming Shan[3], Yixuan Yuan[2‡], Yefeng Zheng[9‡], Lijun Lu[1‡]**

[1]Southern Medical University, [2]The Chinese University of Hong Kong, [3]Fudan University, [4]Northwestern Polytechnical University, [5]Xi'an Jiaotong University, [6]Institute of Automation, Chinese Academy of Sciences, [7]University of British Columbia, [8]Microsoft Research, [9]Westlake University

[*]**These authors contributed equally to this work.** [†] **Project Leader.** [‡] **Corresponding authors.**

GitHub Repo



## Abstract

Advances in multimodal large language models (MLLMs) are extending radiological artificial intelligence (AI) beyond task-specific image analysis toward multimodal understanding and reasoning. Volumetric radiology, however, presents a fundamental representational mismatch: clinical interpretation often requires the integration of full-volume spatial context with acquisition-dependent quantitative information, whereas current MLLMs are commonly conditioned on selected two-dimensional (2D) images, compressed visual representations, or report-derived text. As a result, reliable volumetric radiology AI requires both representations that preserve task-relevant three-dimensional (3D) information and systems that can access, verify, and integrate this information during image interpretation and across clinical workflows.

In this Review, we examine more than 200 publications available through July 2026. Specifically, we organize the literature around volumetric representation and multimodal understanding at the model level, agentic orchestration at the system level, and relate both levels to clinical applications and their evaluation. Within this structure, we first review how volumetric foundation models learn 3D representations, align them with clinical language, and compress them for language-model conditioning. We then assess how agentic systems extend MLLMs beyond single-pass inference through planning, specialized tools, memory, and workflow interaction. Together, these analyses distinguish settings in which selected 2D views or report-mediated reasoning can suffice from those that warrant native volumetric modeling. To support critical assessment across these settings, we introduce a Claim–Design–Validation framework to assess whether technical, workflow, and clinical claims are supported by an appropriate design and a validation setting commensurate with the intended use. Across the reviewed literature, the need for native volumetric modeling and agentic capabilities depends on the spatial, quantitative, contextual, and workflow requirements of the intended task rather than on architectural complexity alone. Accordingly, progress toward clinically credible volumetric radiology AI will require faithful volumetric representation, traceable system behavior, validation commensurate with the intended claim, and clearly defined human oversight in realistic workflows.

# Contents

# 1 Introduction

Volumetric radiology presents a distinctive challenge for medical artificial intelligence (AI) because clinically relevant evidence is distributed across three-dimensional (3D) space, acquisition settings, and longitudinal examinations [1–6]. Computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and related modalities capture anatomical structure, pathological extent, physiological characteristics, treatment response, and longitudinal change. Relevant findings may be small or sparsely distributed and may depend on continuity between slices or relationships between anatomical structures. Their interpretation also requires quantitative measurements, prior examinations, and clinical context [7–9]. The difficulty therefore lies not only in processing large volumetric datasets, but also in maintaining the integrity of the spatial and contextual evidence required for a particular clinical task.

Earlier radiology AI approaches, including classical image-processing, machine-learning, and subsequent deep-learning methods, largely addressed this complexity through task-specific solutions for segmentation, detection, classification, measurement, and report generation [10–13]. These systems achieved substantial progress on well-defined tasks, but their representations and outputs were typically optimized for a fixed prediction objective. Foundation models broaden this paradigm by learning reusable representations that can be adapted across datasets, anatomical regions, and downstream applications. Volumetric self-supervised approaches, including Models Genesis [14], Swin UNETR [15], VoCo [16], and related methods [6, 17, 18], learn transferable anatomical and spatial priors directly from 3D medical images. 3D vision-language pre-training approaches, including CT-CLIP [19], Percival [20], and other related models, further align volumetric representations with radiological reports and clinical semantics. Together, these developments establish a representational basis for radiology systems that can generalize beyond a single task or label space.

Against this background, radiology is entering the era of multimodal large language models (MLLMs), which are rapidly advancing in multimodal interpretation, clinical reasoning, and tool use [1, 3, 21, 22]. However, volumetric imaging highlights a key constraint: radiological evidence is often distributed across the full 3D examination, whereas most MLLMs [5, 23–27] operate on selected two-dimensional (2D) images, compressed representations, or report-derived text. Such inputs may suffice when decisive evidence is confined to selected images or already captured in reports. By contrast, tasks involving lesion extent, anatomical relationships, quantitative burden, phase- or sequence-specific characteristics, or longitudinal change require integration across the volume. Accordingly, the need for native volumetric modeling should be determined by the evidence requirements of the task, specifically whether the available representation preserves the spatial and clinical information necessary to support reliable outputs. These evidence requirements are driving the field toward native 3D modeling, which becomes essential when clinically relevant information is distributed across the volume and cannot be reliably preserved by selected 2D representations. Domain-specific volumetric foundation models [3, 28–31] are well aligned with the anatomical and modality-specific structure of radiology data, while MLLMs offer flexible language-based interaction and broader reasoning capabilities. Progress in volumetric radiology AI therefore depends on integrating these complementary strengths.

Agentic systems extend this integration from model-level inference to workflow-level interaction. Direct MLLM inference generally produces an output from a predefined visual input and prompt, whereas an agentic system can progressively acquire and evaluate evidence over multiple steps [32–34]. It may identify relevant series or anatomical regions, invoke specialized tools for image analysis and measurement, retrieve clinical context, compare prior examinations, retain intermediate observations, and revise its conclusions before returning uncertain decisions to human users. Systems such as CT-Agent [35], 3DMedAgent [36], RadAgent [37], CT-Flow [38], and Radiologist Copilot [39] illustrate this workflow-oriented direction. Such capabilities are particularly relevant to volumetric radiology because clinical interpretation is inherently procedural. Radiologists navigate image volumes, compare phases and sequences, perform measurements, review previous examinations, and integrate imaging evidence with clinical context. Agentic systems should therefore be understood not primarily as autonomous replacements for radiologists, but as mechanisms for targeted inspection, tool-mediated analysis, evidence verification, and human-

supervised workflow coordination. Together, volumetric foundation models and agentic systems represent two mutually reinforcing research trajectories: advances in volumetric representation expand the evidence available to agents, while agentic workflows introduce new requirements for representation fidelity, clinical grounding, provenance, and reliability (Figure 1).

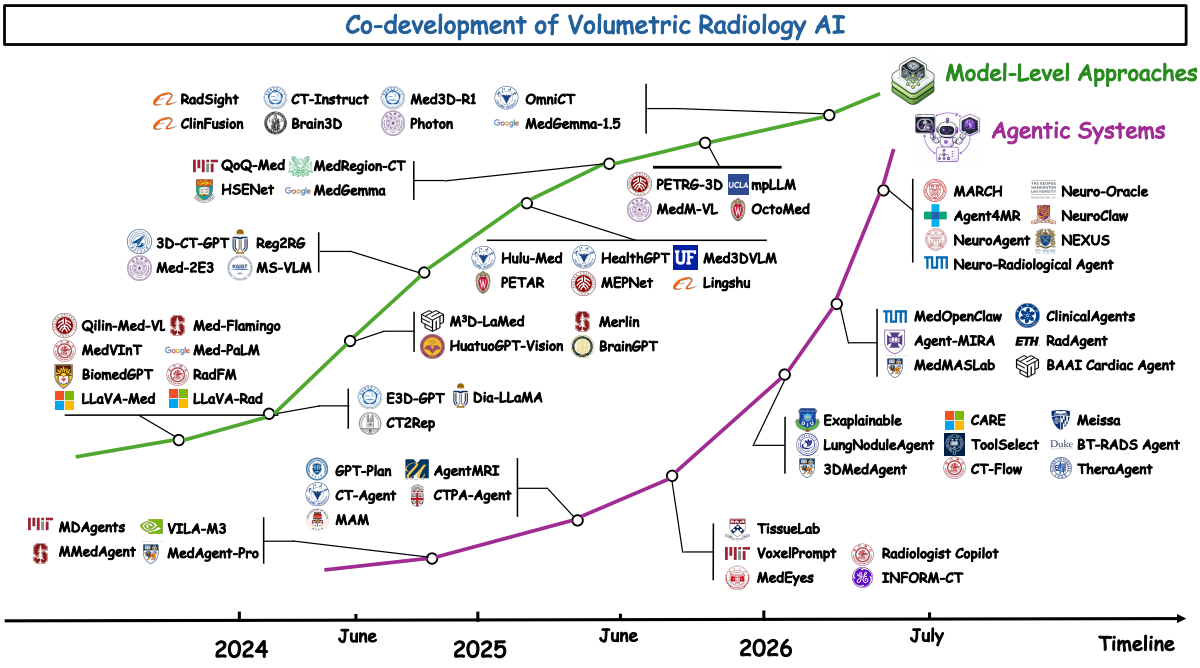


Figure 1: **Co-development of model-level approaches and agentic systems in volumetric radiology AI.** The figure presents two complementary research directions. Model-level approaches, shown in green, include medical MLLMs, volumetric foundation models, and volumetric radiology MLLMs that learn or expose reusable spatial and vision–language representations. Agentic systems, shown in magenta, coordinate these representations through iterative planning, specialized tools, context management, and workflow interaction.

Building on these technical capabilities, their clinical significance ultimately depends on how they are evaluated against the requirements of their intended tasks. Reporting, diagnosis, prognosis, segmentation, measurement, treatment planning, and longitudinal assessment impose different requirements for input completeness, spatial precision, uncertainty management, and human oversight [40, 41]. Evaluation must therefore align with both the evidence requirements and the operational conditions of the intended application. Fluent report generation does not by itself demonstrate faithful image-based interpretation. Likewise, isolated improvements on classification or question-answering benchmarks do not necessarily establish spatial grounding or clinical usefulness. Similarly, successful tool invocation does not demonstrate that an agent selected the appropriate tool, interpreted its output correctly, or improved the surrounding workflow. These considerations make rigorous evaluation central to trustworthy AI, in which technical capability should be accompanied by traceable evidence paths, explicit provenance and audit trails, and clearly defined human responsibility [42]. Clinically credible systems must therefore preserve the volumetric evidence relevant to the task, ground their outputs in imaging and clinical context, and support rather than displace human agency. Their clinical utility should be demonstrated in realistic workflows in which clinicians can examine, challenge, and override AI-supported conclusions while retaining ultimate authority and responsibility for clinical decisions.

Recent reviews have examined generalist medical AI and foundation models [2, 22], medical MLLMs [43, 44], 3D vision-language modeling [1, 45], and agentic radiology [46]. These perspectives, however, are typically examined at different levels of the system and are rarely connected through the question of how volumetric evidence is represented and preserved. In particular, existing surveys have not systematically traced how native volumetric information is transformed into multimodal representations, exposed to language models, retrieved or verified by agents, and ultimately used to support clinical


Figure 2: **Organization of this review.** The review connects model-level volumetric representation and multimodal interpretation with system-level agentic orchestration, and relates both to clinical applications and evaluation. Volumetric foundation models determine what anatomical, quantitative, and clinical information is represented. MLLMs and agentic systems determine how that information is queried, supplemented, verified, and used within clinical and technical workflows. Clinical applications define the intended claims, evidence requirements, and evaluation conditions under which these capabilities should be interpreted.

outputs. As a result, the relationship between representation fidelity, downstream reasoning, workflow interaction, and clinical evaluation remains insufficiently characterized. This review draws on more than 200 publications available through July 2026 and organizes the literature around volumetric foundation models and volumetric radiology MLLMs at the model level, agentic radiology systems at the system level, and clinical applications and evaluation resources. Primary emphasis is placed on work published since 2022, with selected earlier studies included to establish the field's methodological foundations. We place volumetric evidence at the center of this synthesis and examine how its preservation, accessibility, and use are shaped by model design and system behavior. Specifically, we consider when 2D or report-mediated reasoning is sufficient, when native volumetric modeling is necessary, and how language-based and agentic systems can retrieve, verify, and act on spatial evidence while preserving clinical fidelity. We use the term "volumetric radiology AI" to encompass representation learning from 3D medical images, alignment between volumetric and clinical-language information, MLLM-based interpretation, tool-mediated agentic

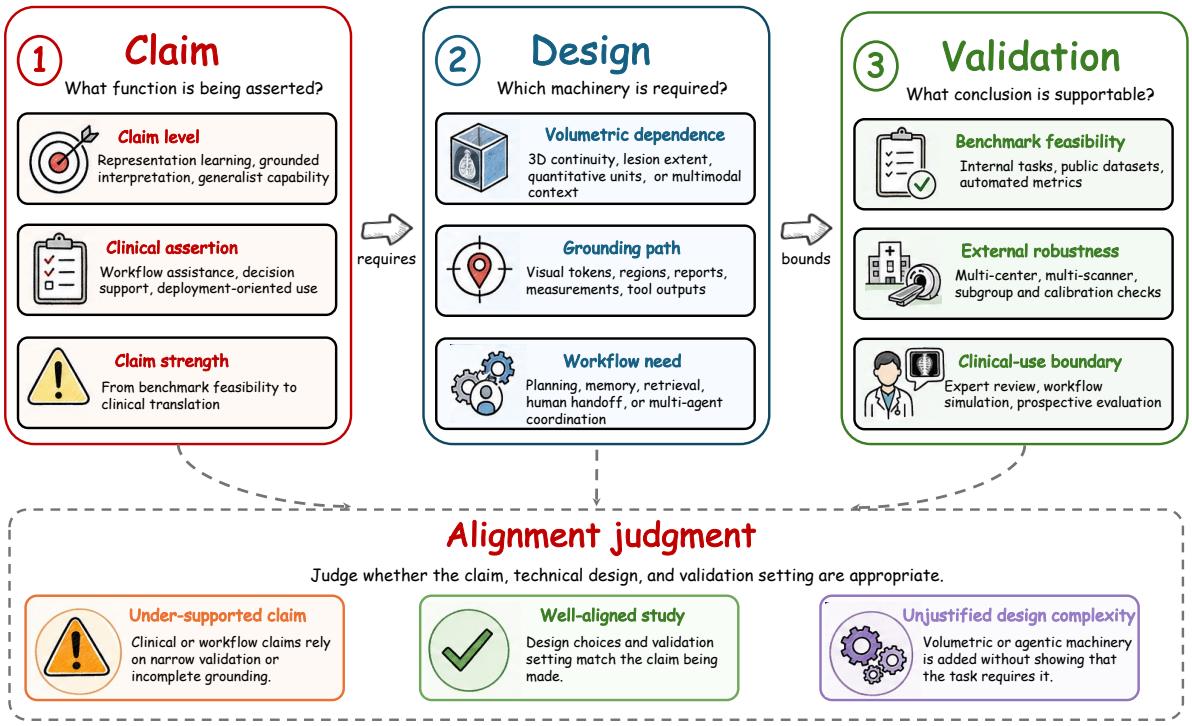


Figure 3: **The Claim–Design–Validation alignment framework.** The framework evaluates the alignment among a study's stated claim, technical design, and validation setting. The claim defines the intended capability or clinical use; the design specifies how the necessary evidence is represented and used; and the validation setting determines the strength and scope of the conclusions that can be supported.

systems, and their evaluation within radiological tasks and workflows. As summarized in Figure 2, we organize the review using a model–system framework linked to clinical applications and evaluation: how volumetric evidence is represented, how it is accessed and acted upon, and how the resulting capabilities are evaluated in clinically relevant settings.

To support a critical rather than purely descriptive synthesis, we use a concise Claim–Design–Validation alignment framework to appraise the studies surveyed in this review (Figure 3). The framework considers whether a study's stated claim is supported by an appropriate technical design and by a validation setting that reflects the intended use. For volumetric radiology AI, this assessment includes whether the task requires native volumetric modeling, multimodal grounding, tool-mediated analysis, memory, planning, or human handoff, and whether the reported evaluation is sufficient to support the corresponding technical, workflow, or clinical claim. The framework therefore complements the evidence-centered perspective of this review by assessing the alignment among task requirements, system capabilities, and validation conditions.

The contribution of this review lies in connecting research that has previously been discussed from largely separate model-level, system-level, and application/evaluation perspectives. Figure 4 summarizes this technical scope and maps the major research directions to the corresponding sections and tables. We first examine volumetric self-supervised representation learning, vision–language alignment, and volumetric radiology MLLMs across 2D-native or selected-view, slice-sequence, hybrid 2D/3D, and native 3D designs. We then consider how agentic systems extend or complement these models through iterative evidence acquisition, specialized tool use, context management, and workflow interaction. Finally, we relate these technical capabilities to clinical applications and evaluation, with particular attention to whether systems preserve task-relevant volumetric and quantitative evidence, ground their outputs in identifiable image and clinical context, maintain traceable provenance, and demonstrate utility under realistic conditions. This perspective provides a structured basis for assessing how model-level representation and system-level orchestration can contribute to clinically credible radiology AI.

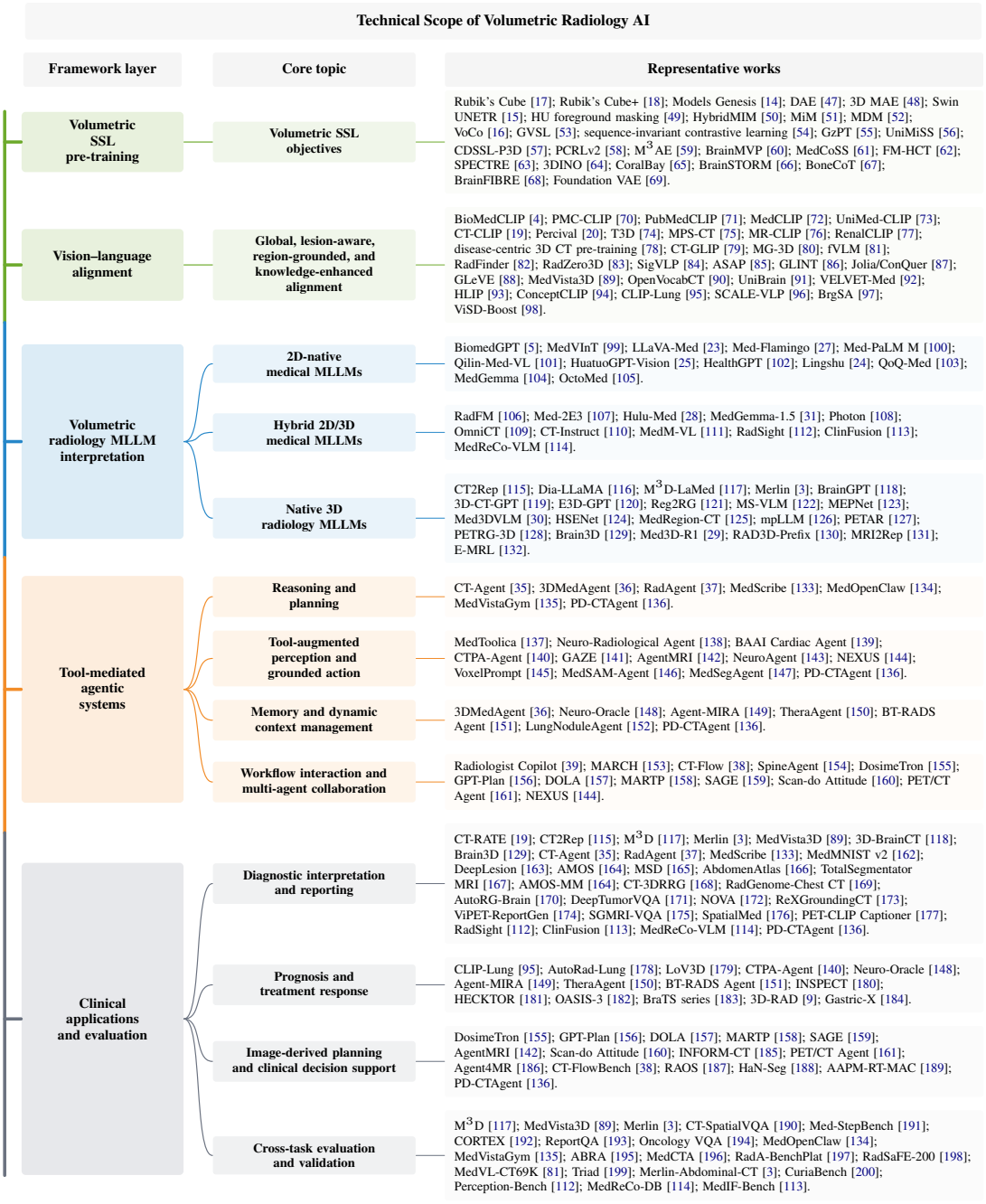


Figure 4: **Technical scope of this review.** The reviewed literature spans volumetric self-supervised pre-training, vision–language alignment, volumetric radiology MLLMs across 2D-native or selected-view, slice-sequence, hybrid 2D/3D, and native 3D designs, agentic radiology systems, and clinical applications and evaluation. Representative methods and resources are organized according to the corresponding sections and tables.

## 2 Preliminaries

This section provides a high-level technical background for the rest of the review. Guided by this field-level model–system–application progression, we trace how volumetric radiology analysis has evolved from task-specific image-analysis pipelines to language-centered multimodal interfaces and, increasingly, to agentic systems that coordinate models, tools, memory, and clinical workflows. The goal is not to duplicate the detailed surveys in later sections, but to establish the conceptual transitions that have driven the emergence of volumetric foundation models and agentic systems.

### 2.1 Traditional Volumetric Radiology Analysis Methods

Early computational paradigm for volumetric radiology treated 3D examinations across diverse radiological and nuclear medicine modalities as structured volumes from which clinically meaningful measurements could be extracted. A typical pipeline consisted of image acquisition and reconstruction, registration or normalization, region-of-interest delineation, handcrafted feature extraction, and a downstream statistical or machine-learning model [201]. Radiomics [10, 11] made this paradigm explicit by converting volumetric images into high-dimensional quantitative descriptors of intensity, shape, texture, and spatial heterogeneity, and by linking these descriptors to diagnosis, prognosis, or treatment response. This represented an important shift from qualitative visual interpretation toward the development of image-derived quantitative biomarkers.

However, traditional feature-engineering pipelines had limited robustness. Their performance depended heavily on segmentation quality, scanner protocol, reconstruction kernel, voxel spacing, feature standardization, and the chosen classifier. Because the representations were usually handcrafted for a specific organ, lesion, or endpoint, transferability across institutions, modalities, and tasks was limited. In addition, most systems produced fixed, task-specific outputs rather than interactive, context-aware support: they could perform classification, segmentation, or retrieval, but they could not naturally discuss uncertainty, compare findings across prior examinations, or adapt their workflow to a clinician's question.

Deep learning changed the representation layer by replacing handcrafted descriptors with learned features. In volumetric radiology, this transition was especially important because isolated 2D slices may fail to capture information contained in the volumetric context. Early 3D convolutional systems such as DeepMedic [12] used multi-scale 3D CNNs to combine local detail with contextual information for brain lesion segmentation. Encoder-decoder architectures then became the dominant template for dense volumetric prediction. 3D U-Net [13] extended the U-Net design to sparse-to-dense volumetric segmentation, while V-Net [202] introduced a fully convolutional 3D architecture and Dice-based optimization for imbalanced medical segmentation. Later, nnU-Net [203] demonstrated that robust biomedical segmentation depends not only on network design but also on systematic self-configuration of preprocessing, architecture, training, and post-processing across datasets. Transformer-based 3D models such as UNETR [204], and Swin UNETR [15] further expanded the receptive field and made long-range anatomical context easier to model.

Despite these advances, supervised 3D deep networks remained largely task-specific. They usually required curated labels, produced outputs within a fixed prediction space, and were optimized for one clinical objective at a time. Their capabilities did not transfer automatically across tasks: a segmentation model did not automatically become a report generator; a classifier did not automatically expose a reasoning trace; and a volumetric backbone trained on one dataset did not necessarily transfer to another modality or institution. This limitation motivates the next stage of the field: foundation-style representation learning, vision-language alignment, and multimodal models that can reuse volumetric perception across tasks.

### 2.2 Large Language Models

Large language models constitute the second background pillar for modern volumetric radiology systems. The Transformer [205] architecture enabled highly parallel sequence processing with attention-based modeling of long-range dependency. GPT-3 [206] further demonstrated that large-scale autoregressive language modeling can support strong in-context learning. Instruction tuning and reinforcement learning

from human feedback, represented by InstructGPT [207], improved the usability of LLMs in interactive settings by aligning next-token prediction with user intent and dialogue behavior. Prompting methods such as chain-of-thought reasoning [208] demonstrated that LLMs can generate explicit intermediate reasoning steps, while ReAct-style prompting [209] connected reasoning with action selection and tool use.

For radiology, the value of LLMs does not lie in replacing image interpretation with text. Rather, their value lies in supporting clinical language understanding, report generation, question answering, explanations of differential diagnoses, knowledge retrieval, and reasoning over contextual information. These capabilities are important because radiology interpretation extends beyond visual pattern recognition. A comprehensive diagnostic process often requires the integration of imaging findings with anatomical knowledge, prior reports, clinical history, guideline thresholds, uncertainty estimation, and downstream clinical decisions [210–213].

At the same time, LLMs expose a key mismatch. Their native input is text, whereas volumetric radiology information is dense, spatial, quantitative, with clinically important evidence often sparsely distributed. If a CT volume is reduced to a few sentences or a small set of slices before reaching the LLM, subtle findings, longitudinal correspondence, and lesion-level information may be lost. Therefore, the role of the LLM in volumetric radiology should be understood as a reasoning and communication layer built upon reliable spatial grounding rather than as a replacement for volumetric perception. This motivates the subsequent discussion of representation learning, vision-language alignment, volumetric radiology MLLM architectures, and agentic radiology workflows.

## 2.3 AI Agents in Healthcare and Radiology

AI agents extend the capabilities of LLMs and MLLMs from answer generators into goal-directed systems. At minimum, an agent maintains a task state, plans intermediate steps, invokes tools, observes results, updates memory, and decides when to answer or defer to human review. This framing is particularly natural in medicine because clinical reasoning is distributed across heterogeneous evidence sources: images, reports, laboratory values, prior examinations, guidelines, measurements, and specialist tools. General medical agents such as MDAgents [32], MMedAgent [33], and MedAgent-Pro [34] have therefore explored adaptive collaboration among LLMs for medical decision-making, tool selection for multimodal medical tasks, and evidence-based diagnostic workflows.

Radiology agents differ from general medical agents because the relevant evidence is spatially structured and image-grounded. A useful radiology agent should not only produce a plausible conclusion, but also inspect the correct image regions, invoke the appropriate measurement or segmentation tools, preserve anatomical provenance, compare with prior examinations when needed, and expose sufficient intermediate observations and tool outputs for clinical review. Early radiology-oriented agents already show this shift. MedRAX [214] integrates chest X-ray analysis tools with an MLLM-based reasoning workflow, while CT-Agent [35] and 3DMedAgent [36] extend agentic reasoning to volumetric CT question answering and 3D medical analysis. RadAgent [37] and CT-Flow [38] further emphasize stepwise CT interpretation and workflow orchestration. Radiologist Copilot [39] and the Neuro-Radiological Agent [138] highlight the role of human-in-the-loop interaction and tool-mediated clinical assistance.

The distinctive value of agentic systems in volumetric radiology lies in controlled access to image-grounded evidence and coordination across workflow steps. When an intended task depends on through-plane continuity, the system must be able to access the corresponding full-volume evidence. Compared with generic clinical agents, radiology agents must link language outputs to image regions, measurements, series, and tool provenance. This review therefore treats agentic radiology as a system-level extension of, or complement to, volumetric foundation models and volumetric radiology MLLMs: representation learning provides perceptual priors, vision–language alignment provides language-addressable semantics, MLLMs provide interactive interpretation, and agentic systems organize these capabilities within traceable, human-supervised workflows.

# 3 Foundation Models in Volumetric Radiology

The rapid advancement of artificial intelligence in medical imaging has ushered in a new era of foundation models for volumetric radiology analysis. In this review, we use foundation models to refer to models pretrained on broad volumetric, visual, textual, or multimodal medical data whose representations or interfaces can be adapted to multiple downstream tasks rather than optimized for a single fixed label space. Under this definition, volumetric self-supervised encoders, vision-language pretrained models, and 3D or hybrid MLLMs are connected by a shared role: they provide reusable perceptual, semantic, or instruction-following capabilities for more than one clinical task. Unlike 2D natural images, 3D radiological data pose unique challenges due to their complex anatomical structures, volumetric nature, and the scarcity of high-quality expert annotations. Before volumetric radiology models can support generation, dialogue, or multi-step understanding, they must first acquire reliable volumetric representations, connect these representations to clinical semantics, and expose the resulting visual information to language models in a form that supports reasoning and communication.

We therefore organize the model landscape as a progression from volumetric self-supervised representation learning to vision-language alignment and then to MLLMs. This progression is functional rather than purely architectural. Volumetric self-supervised learning (SSL) explains how transferable anatomical and spatial priors are learned from scans; vision–language pre-training explains how these priors become addressable by clinical language; and MLLMs explain how visual representations are converted into language-model context for generation, dialogue, and instruction following. The challenge of compressing dense visual information becomes increasingly explicit across this progression. Two-dimensional slice encoders, native 3D encoders, and hybrid 2D/3D encoders appear in SSL, vision–language pre-training, and MLLMs; their consequences become most explicit in MLLMs because the encoded volume must be tokenized, compressed, and exposed to an autoregressive decoder. Figure 5 summarizes this three-stage foundation-model layer and the shared constraint of preserving clinically decisive 3D information for grounded reporting.

## 3.1 Volumetric Self-Supervised Representation Learning

Volumetric self-supervised representation learning provides the perceptual substrate of volumetric radiology foundation models. Its scope is narrower than the full representation-learning literature: supervised, semi-supervised, and weakly supervised training remain important for task-specific radiology systems, but they usually depend on predefined labels, organs, lesions, or clinical endpoints. In contrast, SSL is particularly suitable for the foundation-model layer because unlabeled 3D scans are far more scalable than dense voxel-level annotations or carefully paired reports, while the volumes themselves contain rich intrinsic structure, including slice continuity, anatomical topology, modality-specific appearance, and spatial redundancy. By converting these structures into pretext supervision, SSL can learn transferable anatomical and spatial priors before the model is specialized for language alignment, instruction following, or downstream clinical tasks. Accordingly, this subsection focuses on volumetric SSL objectives, including masked modeling, context restoration, deformation prediction, and contrastive learning, and on how these objectives support reusable 3D visual backbones for later vision–language pre-training and MLLM integration.

### 3.1.1 Self-Supervised Objectives

SSL aims to learn transferable anatomical representations from unlabeled radiology volumes by optimizing pretext tasks that exploit the inherent structural redundancy of volumetric data. Formally, given an unlabeled dataset $\mathcal{D} = \{\mathbf{X}_i\}_{i=1}^{N}$, SSL optimizes an encoder $f_\theta$ by minimizing a pretext loss:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{\mathbf{X}\sim\mathcal{D}}[\mathcal{L}_{\text{pretext}}(\mathbf{X}, \theta)]. \tag{1}$$

In the context of 3D medical imaging, three broad objective families are especially influential:
**Masked Image Modeling (MIM).** Extending the success of masked autoencoders (MAE [215]) to 3D, this approach minimizes a reconstruction loss:

$$\mathcal{L}_{\text{recon}} = \|g_\phi(f_\theta(\tilde{\mathbf{X}})) - \mathbf{X}\|_2^2, \tag{2}$$

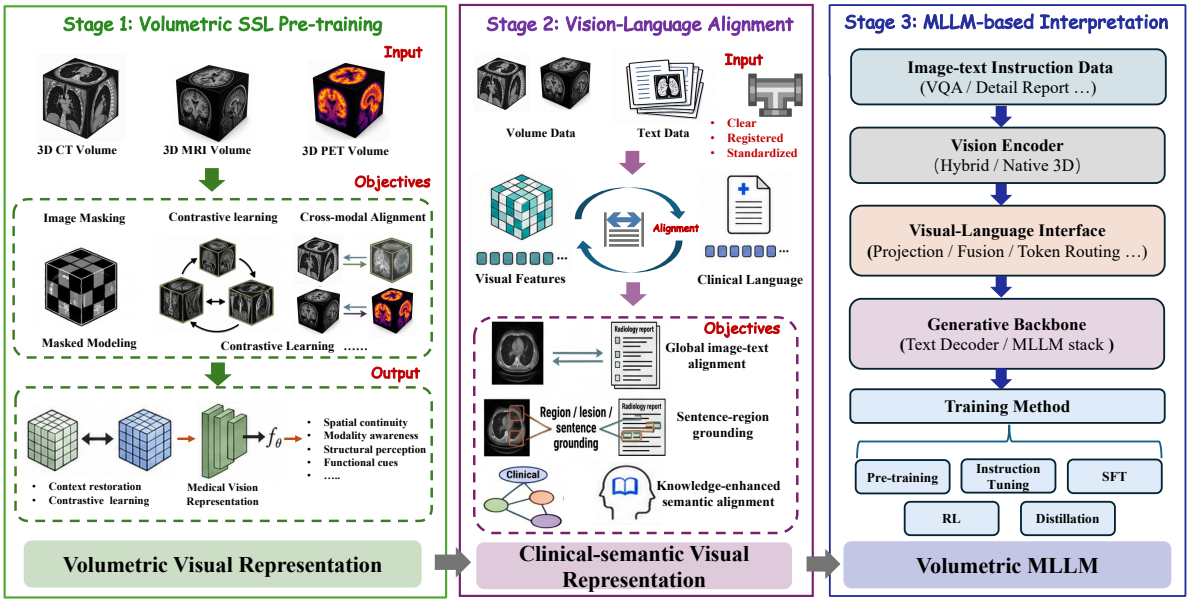


Figure 5: **Model-level foundation and MLLM components for volumetric radiology.** The model landscape is organized into three linked functional stages. Volumetric self-supervised representation learning acquires anatomy-aware priors from unlabeled scans through masked, restorative, deformation-based, and contrastive objectives. Vision–language alignment makes these representations addressable through clinical language at global, lesion-aware, region-grounded, and knowledge-enhanced levels. MLLM-based interpretation couples vision encoders, vision–language interfaces, generative backbones, and training curricula.

where $\tilde{\mathbf{X}}$ is a masked version of the volume. By forcing the model to recover missing voxels, 3D MIM encourages long-range spatial modeling across the volume.

**Contrastive Learning.** This paradigm minimizes an InfoNCE [216] loss:

$$\mathcal{L}_{\text{nce}} = -\log \frac{\exp(\text{sim}(z_i, z_i^+)/\tau)}{\sum_j \exp(\text{sim}(z_i, z_j)/\tau)}, \tag{3}$$

to maximize the similarity between differently augmented views of the same instance. In general, contrastive learning encourages the model to learn discriminative, high-level semantic features by distinguishing between diverse anatomical structures in the latent space. In radiology, however, the design of positives and negatives must be handled carefully because different crops may share identical tissues, and clinically relevant abnormalities may be spatially sparse.

Together, these SSL frameworks establish the visual perception required for subsequent multimodal integration, allowing a 3D encoder to acquire anatomical and spatial priors before linguistic supervision.

### 3.1.2 Representative Approaches

For clarity, we group the representative approaches according to the dominant pre-training signal emphasized in each method. In the surveyed literature, these signals commonly arise from restoring disrupted anatomy, reconstructing masked or deformed structures, enforcing geometric or contrastive invariance, or scaling SSL across dimensions, modalities, and datasets.

**Heuristic Context Restoration.** Early frameworks constructed self-supervision through handcrafted image transformations. For instance, Models Genesis [14] systematically designed transformation strategies such as non-linear appearance mapping, local pixel shuffling, and out-painting, forcing the network to restore the original anatomical structure. Recent extensions, such as DAE [47], combine local channel masking with low-level feature perturbations (e.g., adding noise and downsampling) to increase reconstruction difficulty and encourage recovery of granular anatomical details.

**Anatomically-Guided and Hierarchical Masked Image Modeling.** The MIM paradigm has been widely adapted for 3D representation learning, with initial transformer-based approaches (e.g., 3D MAE [48] and

Swin UNETR [15]) proving its strong transferability. However, naive random masking is inefficient when large portions of 3D scans contain air or background. To address this, HU-based foreground masking [49] leverages Hounsfield Unit density distributions to specifically mask and reconstruct diagnostically meaningful tissue regions. To handle the high dimensionality and varied structural scales of 3D data, hierarchical designs such as HybridMIM [50] and MiM [51] reconstruct features at both coarse regional levels and fine pixel levels simultaneously. Moreover, Masked Deformation Modeling (MDM [52]) incorporates spatial transformations by predicting dense deformation fields alongside voxel intensities, thereby introducing registration-aligned structural priors into the MIM process.

**Geometry-Aware and Invariant Contrastive Learning.** Standard contrastive methods often construct false negative pairs in medical volumes because different cropped patches may share identical semantic tissues or backgrounds. To resolve this, researchers inject geometric and topological priors into the contrastive objective. The VoCo framework [16] extracts base crops and predicts the contextual position of random crops by contrasting their similarities, implicitly encoding organ layout relations into the representations. Similarly, Geometric visual similarity learning [53] enforces local and global semantic matching guided by topological invariance. Furthermore, to ensure invariant representations across varied acquisition protocols, sequence-invariant contrastive paradigms align augmented contrasts derived from quantitative MRI [54], while human-prior aligned methods (e.g., GzPT [55]) leverage eye-tracking sequences to bias contrastive objectives toward clinically relevant regions.

**Cross-Dimensional and Foundation-Scale Self-Supervised Learning.** To reduce the gap between abundant 2D slices and scarcer full-volume supervision, cross-dimensional frameworks such as UniMiSS [56] and CDSSL-P3D [57] jointly exploit planar and volumetric data. Similarly, to maintain high-level semantics alongside pixel-level precision, multi-task frameworks such as PCRLv2 [58] simultaneously optimize contrastive comparison and multi-scale image restoration. Recent work has also expanded SSL toward foundation-scale and multimodal pre-training. Models such as $M^3$AE [59] and BrainMVP [60] address naturally grouped multi-sequence MRI inputs, supporting cross-sequence representations even when some sequences are missing. MedCoSS [61] further broadens this direction by formulating multi-modal medical SSL as a continual pre-training problem across clinical reports, X-rays, CT, MRI, and pathological images, using rehearsal-based learning to mitigate modality conflicts and catastrophic forgetting. Meanwhile, large-scale CT/MRI foundation models (e.g., BrainSTORM [66], FM-HCT [62], SPECTRE [63], and 3DINO [64]) use broader pre-training corpora and larger compute budgets to improve transfer across anatomical regions, scanners, and tasks. Recent CT-specific SSL work further emphasizes this scaling direction: CoralBay [65] extends DINO-style self-distillation to 3D CT with a hierarchical Swin transformer and multi-scale feature distillation, showing how foundation-scale SSL can preserve both global semantics and local anatomical structure.

In summary, volumetric SSL mitigates the annotation bottleneck by converting intrinsic 3D structure into transferable supervision. By encoding anatomical continuity, topological invariants, modality-specific appearance, and spatial relationships, these pre-training objectives provide reusable visual backbones for later alignment and MLLM integration.

## 3.2 Vision–Language Alignment

While volumetric SSL establishes robust anatomical perception, it remains inherently language-agnostic. To make radiological representations addressable through clinical semantics and diagnostic language, vision–language pre-training, largely inspired by CLIP[217], aligns visual features with paired clinical text. This stage should be understood as a semantic bridge rather than simply another generic training strategy: it converts visual features into a language-addressable representation space for retrieval, open-vocabulary recognition, report grounding, and downstream MLLM integration.

### 3.2.1 Alignment Objectives

The most influential formulation of vision–language alignment is the CLIP-style contrastive objective [217], which can be viewed as a cross-modal extension of the InfoNCE principle [216]. In visual contrastive learning, two augmented views of the same image or volume form a positive pair, while other samples serve as negatives. Vision–language contrastive learning instead aligns an image or volume with

its associated report and separates mismatched pairs. In radiology, this enables the visual encoder to learn clinically meaningful representations organized around anatomy, findings, diseases, and diagnostic impressions.

Given a dataset $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^{N}$, where $\mathbf{X}_i$ is the $i$-th image or volume, $\mathbf{T}_i$ is its associated report, and $N$ is the number of pairs, vision–language pre-training jointly optimizes a vision encoder $f_\theta$ and a text encoder $h_\omega$. For a mini-batch of $B$ pairs, the encoders produce $\ell_2$-normalized embeddings:

$$z_i^v = \frac{f_\theta(\mathbf{X}_i)}{\|f_\theta(\mathbf{X}_i)\|_2}, \qquad z_i^t = \frac{h_\omega(\mathbf{T}_i)}{\|h_\omega(\mathbf{T}_i)\|_2}. \tag{4}$$

The embeddings $z_i^v$ and $z_i^t$ represent the visual study and report, respectively, while $\theta$ and $\omega$ denote the encoder parameters.

The image-to-text contrastive loss is

$$\mathcal{L}_{v\to t} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\mathrm{sim}(z_i^v, z_i^t)/\tau\right)}{\sum_{j=1}^{B} \exp\left(\mathrm{sim}(z_i^v, z_j^t)/\tau\right)}. \tag{5}$$

In this expression, $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, $\tau$ is the temperature parameter, and $j$ indexes the candidate reports in the mini-batch.

Symmetrically, the text-to-image loss is

$$\mathcal{L}_{t\to v} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\mathrm{sim}(z_i^t, z_i^v)/\tau\right)}{\sum_{j=1}^{B} \exp\left(\mathrm{sim}(z_i^t, z_j^v)/\tau\right)}. \tag{6}$$

This direction treats each report as a query and retrieves its paired visual study from the $B$ candidates.

The final CLIP-style objective averages the two directional losses:

$$\mathcal{L}_{\mathrm{vlp}} = \frac{1}{2}\left(\mathcal{L}_{v\to t} + \mathcal{L}_{t\to v}\right), \tag{7}$$

where $\mathcal{L}_{\mathrm{vlp}}$ denotes the overall bidirectional vision–language alignment objective.

Early medical vision–language pre-training methods largely followed this recipe: they built biomedical image-text pairs from publications, radiology images, or paired clinical captions, and optimized the symmetric contrastive loss so that images and reports became mutually retrievable. The resulting models inherit the practical advantages of CLIP [217]: once image and text share a representation space, disease names or report phrases can be used as textual prompts for zero-shot recognition, retrieval, and downstream adaptation. In medical imaging, however, paired data scarcity, report complexity, and the dense spatial structure of 3D scans make this objective only a starting point. Research in this domain has progressed from 2D foundations to more structured and volumetric alignment strategies, as summarized below.

### 3.2.2 Representative Approaches

To structure this literature, we organize vision–language pre-training approaches according to the level at which image–text correspondence is modeled. Large-scale global alignment methods establish study-level compatibility between volumes and reports; fine-grained and lesion-aware methods introduce local or region-level constraints to preserve spatially sparse findings; and knowledge-enhanced or hierarchical methods further structure the text side so that clinical concepts, negation, uncertainty, synonyms, and report hierarchy are handled more explicitly. These categories are not mutually exclusive, but they reflect increasingly explicit modeling of both spatial correspondence and clinical semantics.

**From 2D Foundations to 3D Global Alignment.** The fundamental challenge of medical vision–language pre-training is the domain gap between generic visual concepts and domain-specific clinical semantics, compounded by data scarcity. Early progress focused predominantly on 2D imaging, where large-scale image-text resources are more accessible. Foundation models such as BioMedCLIP [4], PMC-CLIP [70], and PubMedCLIP [71] achieved robust 2D alignment by leveraging millions of image-caption

pairs extracted from biomedical literature. Concurrently, frameworks such as MedCLIP [72] reduced dependence on paired data by decoupling images and texts, while UniMed-CLIP [73] unified 2D medical-image pre-training.

Translating this progress to volumetric radiology introduces the challenge of pairing dense, high-dimensional volumetric data with corresponding clinical reports. Representative 3D methods first scaled global volume-report alignment. Models such as CT-CLIP [19], Percival [20], and T3D [74] assemble large-scale 3D volume-report corpora (ranging from 25k to 400k pairs) to align global 3D embeddings with corresponding findings. Addressing the scarcity of expert-written reports, MPS-CT [75] extracts low-cost "silver-standard" labels via LLMs to augment pre-training, while MR-CLIP [76] bypasses textual reports entirely by aligning brain MRIs with structured Digital Imaging and Communications in Medicine (DICOM) [1] metadata. Disease-centered vision–language pre-training is also emerging as a complementary route. RenalCLIP [77] narrows CT-language alignment to precision oncology in kidney cancer, while disease-centric 3D CT pre-training with hybrid visual encoding [78] uses disease-level supervision to reduce the semantic coarseness of study-level report matching.

**Fine-Grained and Lesion-Aware Alignment.** A critical limitation of global alignment is information compression. Pooling an entire high-resolution 3D volume into a single embedding can dilute spatially sparse but clinically important findings, particularly small lesions. To mitigate this limitation, recent methods decompose visual and textual inputs to enforce anatomically grounded, fine-grained correspondences.

This decomposition also helps address the "false negative collision" problem prevalent in contrastive batches, where normal tissues or synonymous diseases may be incorrectly penalized; fVLM [81] addresses this issue with anatomy-level CT–report alignment and dual false-negative reduction during contrastive pre-training. For stricter spatial localization, RadFinder [82] leverages text-mined slice indices as weak axial-depth anchors, while RadZero3D [83] and SigVLP [84] adapt 3D chunking architectures with direct patch-to-text or chunk-level cross-attention. Recent methods make this decomposition more explicit. ASAP [85] introduces anatomy-aware semantically adaptive pre-training for volumetric scans, GLINT [86] uses sparse language-image gates to activate query-relevant image patches, and Jolia/ConQuer [87] augments global CT-report contrastive learning with concept-specific queries. Lesion-level grounding methods such as GLeVE [88] further treat report descriptions as atomic semantic units and verify one-to-one correspondence with 3D lesion proposals. To balance details with holistic context, multi-scale strategies such as MedVista3D [89] and OpenVocabCT [90] combine local constraints with global alignment for open-vocabulary segmentation and anomaly detection.

**Knowledge-Enhanced and Hierarchical Semantic Alignment.** Another limitation of standard CLIP is its treatment of clinical text as a flat sequence. Clinical narratives are noisy, structured, and less spatially redundant than volumetric pixels. To reduce misalignment and false positive/negative noise propagated by standard InfoNCE, researchers have added explicit clinical-semantics modules to vision–language pre-training.

To address the rigid, linear processing of standard CLIP, recent frameworks emphasize modeling the inherent structural hierarchies of clinical data. Because standard contrastive learning ignores nested information, methods such as UniBrain [91] and VELVET-Med [92] explicitly decouple contrastive losses across multiple semantic levels, ranging from words and sentences to modality-specific sub-reports and global conclusions, to preserve both granularity and holistic context. Extending this to operational workflows, HLIP [93] implements a slice-to-scan-to-study hierarchical attention mechanism, enabling robust alignment directly on uncurated clinical archives.

Simultaneously, raw text matching is poorly suited to negations, uncertainties, and synonymous terminologies (e.g., "consolidation" vs. "infiltrate"). A prominent strategy is therefore to inject external medical knowledge into the alignment objective. Building on 2D predecessors such as ConceptCLIP [94] and CLIP-Lung [95], 3D frameworks such as SCALE-VLP [96] replace rigid binary matching with a soft-weighted contrastive objective that combines volumetric spatial-coherence weights with report-level

[1] DICOM is the international standard for transmitting, storing, retrieving, processing, and displaying medical imaging information. The current DICOM standard is managed by the Medical Imaging & Technology Alliance, a division of the National Electrical Manufacturers Association (NEMA).

medical-knowledge embeddings derived from a frozen medical language model. This replaces the rigid binary matching of standard InfoNCE with a clinically structured similarity target. Furthermore, to bridge the persistent modality gap caused by disparate information densities between sparse text and redundant volumes, models employ explicit intermediaries. BrgSA [97] introduces an LLM-summarized cross-modal knowledge bank as a shared semantic anchor, while ViSD-Boost [98] addresses the gap by modeling normal-appearance distributions via a vector-quantized variational autoencoder (VQ-VAE), deliberately amplifying pathological deviations before projecting them into the cross-modal space.

In summary, vision–language pre-training bridges the semantic gap between volumetric perception and clinical language. The field is moving from global volume-report matching toward alignment mechanisms that preserve disease concepts, anatomical regions, lesions, and report hierarchy. These text-aligned feature spaces provide the semantic substrate for later MLLMs to support language-mediated interpretation.

### 3.3 Multimodal Large Language Models

Building on SSL backbones and vision–language pre-training, volumetric radiology MLLMs make volumetric representations available for interactive interpretation. The goal is no longer limited to retrieval or classification in an embedding space; it extends to report generation, multi-turn question answering, abnormality localization, uncertainty communication, and instruction following. We therefore treat this layer as an integration problem. The central design question is how a CT, MRI, or PET study is transformed from a dense volumetric representation into visual tokens or cross-attended states that a generative backbone can condition on. This question has four linked components: the vision encoder defines the dimensionality and content of the visual tokens; the vision–language interface projects, resamples, fuses, or routes those tokens; the generative backbone supplies language generation, instruction following, and context handling; and the training strategy determines which parts of the pipeline are pretrained, frozen, aligned, or task-optimized.

To orient the discussion, Table 1 summarizes representative medical MLLMs by input dimensionality, visual backbone, generative backbone or base model, parameter scale, training recipe, and whether an explicit vision-language projector is trained. The table is limited to MLLM systems. It uses input dimensionality to compare how images or volumes are presented to the language model, not to provide a complete taxonomy of SSL or vision–language pre-training encoders. The comparison highlights two broad trends. First, explicit interface learning becomes increasingly important once full volumetric information must be compressed into an LLM-compatible visual token stream. Second, native 3D systems remain more modality-specific than many 2D medical MLLMs because CT, MRI, and PET studies impose heavier demands on token budget, spatial grounding, and instruction data.

Table 1: Representative medical MLLMs relevant to volumetric radiology, grouped by visual architecture. "Visual input" reports the representation presented to the vision pathway after stated preprocessing; a slice sequence is an ordered stack of 2D slices rather than a native 3D tensor. PT: pre-training; SFT: supervised fine-tuning; RL: reinforcement learning; Distill: knowledge distillation. In "Proj.", ✓, ✗, and ∼ denote an explicit, absent, or unspecified visual–language interface, respectively; "–" denotes unavailable or not applicable.

| Model | Visual input | Vision encoder | Gen. backbone | Params | Training | Proj. | Imaging scope |
|---|---|---|---|---|---|---|---|
| *2D-native medical MLLMs* | | | | | | | |
| BiomedGPT [5] | 2D image | VQGAN | BERT | 33M/93M 182M | PT, SFT | ✗ | General medical |
| MedVInT [99] | 2D image | ResNet-50 | PMC-LLaMA / PMC-LLaMA-ENC | 7B | PT, SFT | ✓ | General medical |
| LLaVA-Med [23] | 2D image | CLIP ViT-L/14 | LLaVA | 7B/13B | PT, SFT | ✓ | General medical |
| Med-Flamingo [27] | 2D image | CLIP ViT-L/14 | OpenFlamingo | 8.3B | PT, SFT | ✓ | General medical |
| Med-PaLM M [100] | 2D image | ViT-4B/22B | PaLM-E | 12B/84B 562B | SFT | ∼ | General medical |
| Qilin-Med-VL [101] | 2D image | CLIP ViT-L/14-336 | Chinese-LLaMA2-13B-Chat | 13B | PT, SFT | ✓ | General medical |
| HuatuoGPT-Vision [25] | 2D image | CLIP-Large-336 | LLaVA-v1.5 recipe / Yi-1.5-34B | 8B/34B | PT, SFT | ✓ | General medical |
| HealthGPT [102] | 2D image | CLIP-L/14 | Phi-3-mini / Phi-4 | 3.8B/14B | PT, SFT | ✓ | General medical |
| Lingshu [24] | 2D image | Qwen2.5-VL ViT | Qwen2.5-VL-7B/32B-Instruct | 7B/32B | PT, SFT, RL | ✓ | General medical |
| QoQ-Med [103] | 2D image | Qwen2.5-VL ViT + ECG-JEPA | Qwen2.5-VL | 7B/32B | SFT, RL | ✓ | General medical + ECG |
| MedGemma [104] | 2D image | MedSigLIP | Gemma 3 | 4B/27B | PT, RL, Distill | ✓ | General medical |
| OctoMed [105] | 2D image | Qwen2.5-VL ViT | Qwen2.5-VL-7B-Instruct | 7B | SFT | ✓ | General medical |
| *Hybrid 2D/3D architectures* | | | | | | | |
| RadFM [106] | 2D image / 3D volume | 3D ViT + Perceiver | MedLLaMA-13B | 14B | PT, SFT | ✓ | General radiology |
| Med-2E3 [107] | 3D volume | M3D-CLIP + SigLIP | Phi-3-mini | 3.8B | PT, SFT | ✓ | CT |
| Hulu-Med [28] | 2D image / slice sequence | SigLIP-NaViT | Qwen2.5/Qwen3 Instruct | 4B/7B/8B 14B/32B | PT, SFT | ✓ | General medical |
| MedGemma-1.5 [31] | 2D image / slice sequence | MedSigLIP | MedGemma / Gemma 3 | 4B | PT, SFT, RL, Distill | ✓ | General medical |
| Photon [108] | 2D image / 3D volume | 3D ViT + Qwen2.5-VL ViT | Qwen2.5-VL | 3B/7B | PT, SFT | ✓ | CT |
| OmniCT [109] | 2D image / 3D volume | SigLIP + SCE/OSE modules | Qwen2.5 | 3B/7B | PT, SFT | ✓ | CT |
| CT-Instruct [110] | 3D volume | Hybrid ResNet-ViT | Qwen2-VL-7B-Instruct | 7B | PT, SFT | ✓ | CT |
| UniReason-Med [218] | 2D image / slice sequence | frozen Qwen2.5-VL vision tower | Qwen2.5-VL-7B-Instruct | 7B | SFT, RL | ✓ | General medical + CT |
| MedM-VL [111] | 2D image / 3D volume | SigLIP / M3D-CLIP | Qwen2.5-3B-Instruct | 3B | PT, SFT | ✓ | General medical + CT |
| RadSight [112] | 2D image / 3D volume | SigLIP-NaViT + Radar (PlainConvUNet) | Qwen3-VL | 4B/8B | PT, SFT | ✓ | General medical + CT |
| ClinFusion [113] | 2D image / 3D volume | Qwen ViT + DINOv2 + ConvNeXt + PE-3D | Qwen3-VL | 8B/32B | PT, SFT | ✓ | General medical |
| MedReCo-VLM [114] | 2D/3D image pair | modality-aware MoE ViT | Qwen2.5-7B-Instruct | 7B | PT, SFT | ✓ | General radiology |

Table 1: Representative medical MLLMs relevant to volumetric radiology (continued).

| Model | Visual input | Vision encoder | Gen. backbone | Params | Training | Proj. | Imaging scope |
|---|---|---|---|---|---|---|---|
| *Native 3D radiology MLLMs* | | | | | | | |
| CT2Rep [115] | 3D volume | 3D ViT | – | – | SFT | ✗ | CT |
| Dia-LLaMA [116] | 3D volume | 3D ViT + Perceiver | LLaMA2-7B | 7B | SFT | ✓ | CT |
| $M^3D$-LaMed [117] | 3D volume | 3D ViT + Perceiver | LLaMA-2-7B | 6.7B | PT, SFT | ✓ | General radiology |
| Merlin [3] | 3D volume | I3D ResNet152 | RadLlama-7B | 7B | PT, SFT | ✓ | CT |
| BrainGPT [118] | 3D volume | frozen CLIP ViT-L/14 + Perceiver Resampler | Otter | 7B | SFT | ✓ | Brain CT |
| 3D-CT-GPT [119] | 3D volume | CT-ViT from CT-CLIP | Vicuna-7B | 7B | PT, SFT | ✓ | CT |
| E3D-GPT [120] | 3D volume | 3D MAE ViT-base + 3D convolutional projector | Vicuna-7B | 7B | PT, SFT | ✓ | CT |
| Reg2RG [121] | 3D volume | RadFM 3D ViT + Perceiver + 3-layer ViT3D mask encoder | LLaMA2-7B | 7B | SFT | ✓ | CT |
| MS-VLM [122] | 3D volume | DINO ViT-B/16 + Z-former | Vicuna-7B-v1.5 | 7B | PT, SFT | ✓ | CT |
| MEPNet [123] | 3D volume | ResNet101 | LLaMA3-8B | 8B | SFT | ✓ | Brain CT |
| Med3DVLM [30] | 3D volume | DCFormer-S + SigLIP | Qwen2.5-7B-Instruct | 7.6B | PT, SFT | ✓ | General radiology |
| HSENet [124] | 3D volume | 3D-ViT + 2E3-ViT | Phi-4-4B-Instruct | 4B | PT, SFT | ✓ | CT |
| MedRegion-CT [125] | 3D volume | RAD-DINO (ViT-B) + MAIRA-SEG mask extractor | LLaMA3-8B | 8B | PT, SFT | ✓ | CT |
| mpLLM [126] | 3D volume | M3D-CLIP ViT | Phi-3-Mini-4K-Instruct | 3.8B | SFT | ✓ | Brain MRI |
| PETAR [127] | 3D volume | M3D-CLIP ViT with PET/CT/mask projectors | M3D / Phi-3-4B | 4B | PT, SFT | ✓ | PET/CT |
| PETRG-3D [128] | 3D volume | RadFM 3D ViT + Perceiver | Qwen3-8B | 8B | SFT | ✓ | PET/CT |
| Brain3D [129] | 3D volume | 3D-ViT | MedGemma-1.5-4B-IT | 4B | PT, SFT | ✓ | Brain MRI |
| Med3D-R1 [29] | 3D volume | 3D CLIP ViT | Qwen2.5-3B | 3B | PT, SFT, RL | ✓ | CT |
| RAD3D-Prefix [130] | 3D volume | CT-CLIP / CT-ViT + anomaly logits | LLaMA-3.2-1B | 1B | SFT | ✓ | CT |
| MRI2Rep [131] | 3D volume | 3D CNN + visual Transformer | AR Transformer decoder | – | PT, SFT | ✗ | Liver MRI |

### 3.3.1 Vision Encoders

Visual encoder dimensionality is a cross-cutting design axis in volumetric radiology AI. The same 2D, native 3D, and hybrid 2D/3D choices appear in SSL and vision–language pre-training, but their role changes in an MLLM. They determine the visual token sequence or visual states that the language model can condition on. The vision encoder is therefore the point where the representation-learning literature in Sections 3.1 and 3.2 enters the MLLM framework most directly. An LLM can only reason over the visual tokens it receives: if the encoder has not learned volumetric continuity, through-plane relations may be lost; if the encoder has not been aligned with clinical language, the interface must learn the semantic bridge from scarce instruction data. The models in Table 1 can therefore be read as different answers to the same question: should the visual stream inherit a mature 2D biomedical vision–language pre-training prior, a self-supervised 3D anatomical prior, a 3D vision-language prior, or a hybrid combination of these sources?

**2D biomedical and generalist encoders.** Early medical MLLMs mainly adapted the general large vision–language model (LVLM) recipe to biomedical images: a 2D visual encoder, a lightweight connector, and an instruction-tuned language model. BiomedGPT [5] uses a VQGAN-style visual tokenizer and unified image-text token modeling, while MedVInT [99], LLaVA-Med [23], and Med-Flamingo [27] establish the more common pattern of connecting 2D visual features to a medical or general multimodal decoder. These systems are useful baselines because they show how biomedical vision-language priors can support medical VQA, dialogue, and report-style generation with limited architectural change.

For volumetric radiology, however, the 2D-native family is mainly relevant as a source of transferable semantic priors rather than as a sufficient solution. Recent systems based on stronger Qwen-series or MedGemma-style multimodal backbones [104, 219] improve high-resolution image understanding and instruction following, but they still primarily operate on images, selected views, or long slice sequences. Once a CT or MRI study is reduced to a few planar observations, through-plane continuity, lesion extent, and volumetric morphology can be weakened before the vision-language interfaces or LLM receives the visual tokens.

**Native 3D radiology encoders.** Native 3D models accept the volume itself as the primary visual object. This is the most direct trajectory for volumetric radiology tasks in which findings depend on slice-to-slice continuity. CT2Rep [115] is an early representative of this direction: it uses a 3D ViT [220] encoder for CT report generation, avoiding the information loss caused by key-slice selection. Although CT2Rep is not a conversational LLM system, it establishes the important premise that report generation can be conditioned on full-volume representations rather than 2D projections alone.

Several later systems explicitly connect volumetric self-supervised or vision-language priors to generative MLLMs. E3D-GPT [120] first builds a self-supervised 3D foundation model and then uses 3D spatial aggregation before instruction tuning, making it a clear example of SSL-to-MLLM transfer. Dia-LLaMA [116] uses a pre-trained ViT3D and adds disease-aware attention with a prototype memory bank, so its encoder is not merely anatomical but abnormality sensitive. M$^3$D-LaMed [117] constructs large-scale 3D image-text and instruction corpora, couples a 3D vision encoder with spatial token pooling, and extends the visual interface to positioning and segmentation. 3D-CT-GPT [119] uses CT-ViT features derived from CT-CLIP-style pre-training, showing how the global volume-report alignment discussed in Section 3.2 can become the visual front-end of a generative model.

A related group expands the 3D encoder toward broader CT foundation capability. Merlin [3] uses an I3D ResNet-style CT encoder and maps its features to a radiology language model, emphasizing scalability and multi-task competence on CT. CT-CHAT [19] starts from CT-CLIP-style volumetric features and uses attention pooling to build a chat-oriented CT assistant. Med3DVLM [30] combines DCFormer-S with SigLIP-derived semantic signals, so the encoder carries both efficient 3D spatial features and medically aligned visual semantics. HSENet [124] emphasizes local and global spatial information by combining a 3D-ViT branch with a 2D-enhanced branch, which is especially relevant when small findings must be preserved without losing study-level context.

Domain-specialized 3D systems show that the desired encoder depends strongly on modality and anatomy. BrainGPT [118], MEPNet [123], and Brain3D [129] focus on brain CT or MRI, where subtle

structural changes and lesion distributions require a narrower but more specialized visual prior. mpLLM [126] uses M3D-CLIP features for multiparametric brain MRI, so the encoder must handle T1, T1Gd, T2, and FLAIR as complementary sources of volumetric information rather than interchangeable channels. MRI2Rep [131] extends structured report generation to 3D liver MRI, reinforcing that volumetric MLLMs are beginning to specialize beyond chest CT. PETAR [127] and PETRG-3D [128] extend 3D MLLMs to PET/CT, where the visual encoder has to preserve both metabolic uptake and anatomical localization. These models indicate that a generic 3D encoder is unlikely to be sufficient across all clinical settings; the encoder must reflect what the downstream clinical question treats as image-grounded information.

A further trend inside native 3D encoders is the shift from study-level features toward localized representation and spatial grounding. Reg2RG [121] augments volumetric texture encoding with mask-derived geometric information so generated reports can be tied to regions. MS-VLM [122] uses slice-level DINO features and a Z-former to retain cross-slice structure before language generation. MedRegion-CT [125] brings pseudo-mask and region tokens into the visual stream to reduce region-level hallucination, while MedVL-SAM2 [221] links 3D visual encoding with segmentation-oriented prompting. Human-body-prior 3D MLLMs [222] and discriminative-guided spatial grounding [223] pursue the same goal from different angles: the encoder should not collapse the scan into a single global embedding before clinically important locations have been protected. At the efficiency frontier, U-VLM [224] injects hierarchical visual information into multiple decoder layers, and BTB3D [225] redesigns volumetric tokenization itself so fewer but more informative 3D tokens reach the language model.

**Hybrid 2D/3D encoders.** Hybrid encoders occupy a pragmatic middle ground between mature 2D vision–language pre-training systems and native 3D volumetric backbones. One strategy treats the volume as an ordered slice or video sequence. Med-Gemini [226] reuses Gemini's video pathway by interpreting CT depth as a temporal dimension, which lets the model benefit from a powerful pre-existing multimodal stack. Hulu-Med [28] and Fleming-VL [227] similarly exploit strong 2D visual encoders, but add token-reduction and positional strategies so long medical image sequences remain manageable. MedGemma 1.5 [31] extends MedGemma to 3D by representing volumes as long axial slice sequences, showing that a medically aligned 2D encoder can be stretched toward 3D with careful context handling and training.

Efficiency-oriented hybrid methods make the same slice-sequence assumption but focus on selecting the right visual tokens. MedPruner [228] removes redundant slice and patch tokens hierarchically, which is important because adjacent CT slices often contain highly repetitive anatomy. Photon [108] goes further by learning instruction-conditioned token scheduling, so the retained visual tokens can vary with the clinical question rather than following a fixed sampling rule. The adaptation study of Yu et al. [229] reinforces this point from an empirical angle: adapting 2D encoders to 3D is attractive under limited data, but it works best when the model has explicit mechanisms for preserving volumetric consistency. UniReason-Med [218] makes the transfer problem more explicit by aligning 2D images and slice-serialized 3D volumes through a shared grounded reasoning interface with region-token injection.

A second hybrid strategy keeps an explicit 3D branch together with slice-level information. RadFM [106] unifies 2D and 3D scans through visually conditioned generative pre-training, using a 3D-compatible visual stream so both planar and volumetric inputs can be fed to the same MLLM. Med-2E3 [107] combines M3D-CLIP features with SigLIP slice features and uses text-guided inter-slice scoring, allowing the instruction to decide which 2D visual tokens should complement the 3D representation. CTInstruct [110] uses a hybrid ResNet-ViT encoder to support diagnosis, segmentation, report generation, and multiple-choice reasoning with one CT-oriented backbone. MedM-VL [111] systematically compares 2D and 3D CT encoder choices for medical LVLMs, while OmniCT [109] introduces unified slice-volume encoding with spatial consistency and organ-level semantic enhancement. RadSight [112] uses dedicated SigLIP-NaViT and Radar/PlainConvUNet pathways so 2D images and 3D CT volumes retain their native spatial structure before entering a shared Qwen3-VL interface. ClinFusion [113] instead enriches Qwen ViT tokens with DINOv2, ConvNeXt, and PE-3D features through cascaded spatial-aware locality fusion, including a 2D-anchored depth-aware operator for volumetric inputs. MedReCo-VLM [114] combines per-slice and through-depth attention in a modality-aware MoE ViT and applies the shared encoder to paired studies, supporting entity-conditioned comparison across radiography, CT, MRI, and ultrasound. Across these hybrid designs, the encoder balances 2D semantic maturity with 3D anatomical completeness.

Across these families, encoder paradigms differ not only in dimensionality but also in the priors they inherit, the information they preserve, and the bottlenecks they introduce. Table 2 condenses these trade-offs at the family level. Unlike Table 1, which inventories model architectures, this comparison links each encoder paradigm to its main limitation and resulting design implication.

Taken together, these comparisons indicate that no encoder paradigm is uniformly optimal across all tasks. Nevertheless, increasing native volumetric integration raises the ceiling for preserving through-plane and full-volume evidence, and should be prioritized when such evidence is required by the intended clinical claim (Table 2). Three principles emerge. First, tasks that depend on anatomy, spatial continuity, or subtle lesion morphology are likely to benefit from encoders that learn volumetric priors, for example through volumetric SSL or 3D image–text pre-training. Second, tasks involving open-vocabulary recognition, retrieval, report generation, or spatially grounded language additionally benefit from medically aligned semantic priors established through vision–language pre-training. Third, regardless of encoder family, compression should be task-adaptive and should not discard local evidence before it can be integrated with organ-level context and global study semantics. The vision encoder therefore does more than initialize the MLLM pipeline: it sets an upper bound on the visual evidence that subsequent interfaces can make available to the language model.

### 3.3.2 Vision-Language Interfaces: Projection, Resampling, and Fusion

After the vision encoder produces image or volume tokens, the next question is how these tokens are made available to the generative backbone. We use vision–language interface as the umbrella term for modules that project, resample, fuse, prune, route, or inject visual states into the generative context. The term adapter remains useful for linear or multilayer perceptron (MLP) projectors, but it does not fully capture the broader functions of current 3D visual–language interfaces. In volumetric radiology, this interface is the bottleneck between dense volumetric information and the LLM token stream. Its role is broader than dimensional matching: it decides how visual tokens are selected, compressed, ordered, grounded, and inserted into language space. The models in Table 1 therefore expose a design spectrum. At one end, the connector is minimal. BiomedGPT [5] avoids a conventional continuous projector by converting images into discrete VQGAN visual tokens and modeling them together with text tokens. LLaVA-style medical systems such as LLaVA-Med [23], Qilin-Med-VL [101], and R2GenGPT [230] mainly use linear or shallow MLP projectors to map 2D visual features into the LLM embedding space. MedVInT [99]

Table 2: Comparison of visual encoder paradigms for volumetric radiology MLLMs, ordered by increasing native volumetric integration. The rows summarize recurring architectural trade-offs reflected in representative systems rather than a universal performance ranking.

| Encoder paradigm | Dominant prior | Visual evidence emphasized | Characteristic bottleneck | Key design implication |
|---|---|---|---|---|
| **2D-native encoders** | Biomedical image–text alignment and multimodal instruction following | In-plane appearance and medically aligned semantics | No native modeling of through-plane continuity | Natural fit for 2D or selected-view tasks; volumetric use requires explicit cross-slice aggregation |
| **Slice-sequence encoders** | 2D image–language representations extended across ordered slices | Ordered slice content with broader volumetric coverage | Long token sequences; cross-slice structure depends on positional modeling and token reduction | Preserve slice order and select task-relevant tokens before aggressive compression |
| **Hybrid 2D/3D encoders** | 2D semantic features combined with 3D spatial features | Fine in-plane detail and global volumetric context | Explicit fusion and alignment across heterogeneous feature streams | Use task-conditioned fusion when both fine detail and 3D context are required |
| **Native 3D encoders** | Volumetric structure learned through 3D SSL or image–text alignment | Through-plane continuity, lesion extent, and global anatomy | High-dimensional inputs, limited 3D data, and costly token processing | Prioritize when spatial continuity or lesion extent is central; compress without discarding spatial structure |

explicitly compares MLP and transformer-decoder projection variants, while Med-Flamingo [27] follows the Flamingo design with a perceiver resampler and gated cross-attention layers. These 2D designs are best viewed as interface baselines for volumetric radiology: they solve vision-language dimensional alignment, but not the more demanding problem of selecting and preserving clinically relevant 3D information.

For volumetric radiology, this simple projector pattern is often insufficient because the interface must reduce thousands of volumetric tokens without erasing small or spatially sparse findings. One direct solution is query-based compression. RadFM [106] pads 2D inputs into a 3D-compatible representation, uses a shared 3D ViT, and applies a perceiver with learned latent queries so variable-size 2D/3D scans become a fixed visual prefix. Dia-LLaMA [116] similarly projects ViT3D patch features through a perceiver before LLM insertion, but further attaches disease-aware attention and a prototype memory bank so the visual prefix is accompanied by abnormality-sensitive diagnostic prompts. M$^3$D-LaMed [117] makes the compression spatially explicit: 3D ViT outputs are reconstructed into a volume-like grid, pooled in 3D space, and then passed through linear or MLP layers, reducing the visual input from dense volumetric patches to LLM-sized tokens while retaining coarse anatomical layout. BrainGPT [118] adapts the Otter/Flamingo paradigm to brain CT with a trainable Perceiver Resampler and cross-gated attention layers, and CT-CHAT [19] uses attention pooling with learned latent queries plus an MLP multimodal projector to connect CT-CLIP features to the LLM. These systems share the same interface logic: learned queries serve as a controlled interface between arbitrarily long volumetric information and the fixed context budget of an autoregressive decoder.

A second family keeps the connector lightweight but makes the spatial path before or inside the connector more structured. CT2Rep [115] is not an LLM-prefix model and therefore has no explicit vision-language projector; its transformer decoder consumes 3D encoded states directly, and CT2RepLong adds cross-attention over previous volumes and reports for longitudinal fusion. Merlin [3] takes a more conventional MLLM route, mapping I3D ResNet features to RadLlama with a single linear adapter and low-rank adaptation (LoRA) tuning. 3D-CT-GPT [119] similarly uses CT-ViT features, 3D average pooling, and a trainable linear projection, with later experiments comparing simple linear projection and MLP alternatives. RAD3D-Prefix [130] revisits the same adaptation problem by studying how diagnostic priors and scaling choices affect LLM adaptation for 3D CT report generation. E3D-GPT [120] argues that a 3D adapter should not behave like a generic Q-Former: its 3D convolutional perceiver merges and projects features while preserving local spatial neighborhoods. HSENet [124] develops this principle further with twin spatial packers for global and local 3D features; its Voxel2Point cross-attention aggregates high-resolution voxel features into centroid tokens, then uses two-layer MLPs to reach the LLM latent dimension. Brain3D [129] follows a compact prefix strategy for brain MRI, compressing inflated 3D transformer tokens to a small fixed set, projecting them through a two-layer MLP, and controlling visual conditioning with a learnable scalar gate.

Hybrid 2D/3D models turn the interface into a fusion module rather than a single projection. Med-Gemini [226] reuses Gemini's video pathway, relying on the model's native long-context multimodal interface rather than a separately specified visual–language projector. Med-2E3 [107] uses separate 3D and 2D branches with MLP-based connectors, then computes text-guided inter-slice scores so the 2D features most relevant to the instruction enhance the 3D feature sequence. MedM-VL [111] studies the connector choice itself: for 3D inputs encoded slice by slice, it compares cross-attention compression to fixed-length tokens against average pooling, and shows that even a simple linear connector can remain competitive when the encoder and training data are well chosen. CTInstruct [110] uses a lightweight linear bridger from hybrid ResNet-ViT features to the text latent space, but deliberately preserves a dynamic token count determined by volumetric patch decomposition instead of forcing every scan into a fixed perceiver prefix. OmniCT [109] introduces a more explicit slice-volume interface: volumetric slice composition and tri-axial positional encodings build unified slice/volume tokens, while its mixture-of-experts (MoE) Hybrid Projection contains slice-specific, volume-specific, and shared projection matrices. This lets projection behavior learned from 2D slices transfer to 3D volumes while still allowing modality-dependent transformations. RadSight [112] and MedReCo-VLM [114] retain lightweight two-layer MLP projectors, but apply them to modality-specific or paired entity-aware features, respectively. ClinFusion [113] makes fusion itself the interface: its CaSL operator incrementally injects local specialist features into the

language-aligned Qwen ViT stream and extends the same mechanism across volumetric depth.

A third line makes the interface region- or mask-aware. Reg2RG [121] reuses RadFM's ViT3D and Perceiver adapter for texture features, but pairs it with a lightweight ViT3D mask encoder so local texture tokens, geometric mask tokens, and global context can be aligned to region-level report descriptions. MedRegion-CT [125] combines $R^2$ token pooling with a mask-driven visual extractor: pseudo-masks generate mask tokens, spatial tokens, and patient-specific attributes that are injected with global CT tokens to reduce region-level hallucination. MEPNet [123] uses a visual adaptor to map scan features to the LLM space, then adds knowledge-driven joint attention to mine entity-specific visual embeddings and learning-status prompts, making the interface partly semantic and entity balanced. PETAR [127] extends this principle to PET/CT by jointly encoding PET, CT, and lesion masks with modality-specific projectors and mask-conditioned PET embeddings; focal crops and mask-aware visual tokens force the language model to describe the highlighted lesion rather than the entire scan without lesion conditioning. PETRG-3D [128] instead uses dual PET and CT streams initialized from RadFM-style 3D encoders, retrains Perceiver samplers for each modality, projects the compressed tokens linearly into the LLM space, and combines them with hospital-style templates. MedVL-SAM2 [221] illustrates the segmentation-oriented version of the same idea, using an MLP-Mixer projection layer to compress 3D visual tokens and a generated `[SEG]` token to drive prompt-based 3D segmentation.

Recent interfaces are also becoming schedulers and routers. Med3DVLM [30] replaces a plain MLP with a multi-scale MLP-Mixer projector that mixes token and channel dimensions across low- and high-level DCFormer features, improving the transfer of both fine spatial and semantic information. mpLLM [126] designs a prompt-conditioned hierarchical MoE projector for multiparametric brain MRI, with routers over modality-level and token-level projection experts so the connector can choose different transformations for T1, T1Gd, T2, and FLAIR according to the question. Med3D-R1 [29] adds adaptive weighted pooling and a residual anchor mapping module, blending projected image tokens with a fixed text-space anchor through token-specific gates before reinforcement learning encourages clinically consistent reasoning. Efficiency-oriented models reinterpret the interface as token budgeting. Hulu-Med [28] uses a two-layer MLP projector but adds medical-aware token reduction for 3D and video inputs. Fleming-VL [227] combines pixel unshuffle, a two-layer MLP projector, and variable visual position encoding so long multi-image sequences occupy less positional space. MedPruner [228] prunes redundant slice and patch tokens hierarchically, Photon [108] learns instruction-conditioned token scheduling with surrogate gradient propagation so retained tokens vary by question, and BTB3D [225] addresses the same bottleneck at tokenization time by producing compact volumetric tokens before LLM injection. U-VLM [224] provides another route by injecting hierarchical visual information into multiple decoder layers rather than relying on one front-loaded prefix. Across these designs, the interface becomes an active control point: it determines not only computational cost, but also whether the LLM receives global study context, local lesion information, modality-specific cues, and spatial provenance needed for a grounded radiological conclusion.

### 3.3.3 Generative Backbones: Text Decoders and Multimodal Foundation Stacks

After visual states have been encoded and routed through the vision-language interfaces, the next design choice is the generation-side model that receives them. We use the term generative backbone for this receiving component. It may be a standalone text decoder, a report-generation decoder, or an inherited VLM/MLLM stack rather than a text-only LLM. This distinction matters because pretrained multimodal stacks already carry assumptions about visual token format, context handling, and cross-modal conditioning, whereas text-decoder-centric systems depend more heavily on the 3D encoder and interface for visual grounding. We therefore organize this subsection by the type of generative backbone inherited by each system.

**Report-generation and unified token-modeling decoders.** Some systems use dedicated report-generation decoders or unified token-modeling frameworks rather than a conversational, instruction-tuned LLM as the generative backbone. BiomedGPT [5] uses a BERT-like generative modeling framework [231], making it closer to unified biomedical token modeling than to instruction-following dialogue. CT2Rep [115] also follows a report-generation route: its transformer decoder generates CT reports

directly from 3D encoded states rather than from a conversational LLM prefix. These systems remain important because they show that full-volume radiology generation can be formulated before the adoption of current instruction-tuned MLLM stacks. MedVInT [99] is a transitional case, moving from task-specific biomedical decoders toward LLaMA-series medical language backbones [232].

**Text-decoder-centric MLLMs.** A large group of medical MLLMs attaches visual states to a text-oriented instruction model. The LLaMA-series [233–235] and related Vicuna-style derivatives became common targets for early medical and volumetric radiology systems, including Dia-LLaMA [116], M$^3$D-LaMed [117], 3D-CT-GPT [119], E3D-GPT [120], Reg2RG [121], MS-VLM [122], MEPNet [123], and MedRegion-CT [125]. Medical variants of this lineage, such as MedLLaMA in RadFM [106] and RadLlama in Merlin [3], provide domain-tuned language priors and report style, but visual factuality still depends on the encoder-interface pipeline.

More recent text-decoder choices reflect the need for stronger instruction following, lower deployment cost, or domain-specific adaptation. Qwen-series text decoders [219] are used in several hybrid or native 3D systems, including MedM-VL [111], Med3DVLM [30], Hulu-Med [28], OmniCT [109], Med3D-R1 [29], PETRG-3D [128], RadSight [112], ClinFusion [113], and MedReCo-VLM [114]. Phi-series models [236] support compact systems such as HealthGPT [102], Med-2E3 [107], HSENet [124], mpLLM [126], and PETAR [127]. In these text-decoder-centric designs, the generative backbone supplies fluency, instruction following, and medical terminology handling. It does not by itself solve volumetric grounding.

**Pretrained VLM/MLLM foundation stacks.** Another group inherits a pretrained multimodal stack rather than a standalone text decoder. LLaVA-style systems such as LLaVA-Med [23], Qilin-Med-VL [101], and HuatuoGPT-Vision [25] adapt an existing visual-instruction architecture to medical images. Med-Flamingo [27] inherits OpenFlamingo-style gated cross-attention [237], while BrainGPT [118] adapts the Otter/Flamingo paradigm [238] to brain CT report generation. Med-PaLM M [100] represents the PaLM-E route [239], where the base model already contains broad multimodal instruction-tuning priors.

Qwen-VL and MedGemma-style systems illustrate the same issue in newer foundation stacks. Lingshu [24], QoQ-Med [103], OctoMed [105], Photon [108], and CTInstruct [110] build on Qwen-VL or Qwen2.5-VL-style multimodal backbones. MedGemma [104], MedGemma 1.5 [31], and Brain3D [129] similarly inherit medically tuned multimodal or vision-language foundation stacks based on Gemma-family models [240]. These systems should not be described as using only an LLM backbone. They inherit multimodal context interfaces developed mainly for 2D images or video-like inputs, and volumetric radiology adaptation must make volumetric tokens compatible with those inherited assumptions.

Across these choices, parameter scale is insufficient as a standalone indicator of clinical usefulness because the three backbone paradigms inherit different capabilities and depend on different routes for visual conditioning. A more informative comparison asks what the backbone contributes, which volumetric information must be supplied by the encoder-interface pipeline, and how the resulting generation should be evaluated. Table 3 summarizes these dependencies.

Taken together, generative backbones should be assessed as part of the complete encoder–interface–generation pipeline rather than in isolation (Table 3). Report-generation and unified token decoders should be judged by task-specific completeness and spatial fidelity; text decoders by whether fluent output remains grounded in the delivered 3D evidence; and pretrained multimodal stacks by whether their inherited interfaces remain compatible with volumetric tokens. Across all three paradigms, clinical usefulness depends on the alignment of spatial grounding, visual-token delivery, and the training objective. The clinical value of a generation-side model therefore depends on its ability to faithfully express the clinically relevant 3D evidence retained by the encoder-interface pipeline.

### 3.3.4 Training Strategies

Training strategies are discussed here according to the specific learning problem addressed by each stage: acquiring volumetric priors, constructing 3D supervision from reports and task labels, enforcing spatial grounding, managing the visual-token budget, and adapting or optimizing the model under limited medical data. This organization reflects a central constraint of volumetric radiology MLLMs: training must preserve clinically relevant volumetric information while converting it into a compact and clinically

grounded language interface.

**From generic alignment to volumetric curricula.** The training strategy of a volumetric radiology MLLM should not be understood as a simple extension of the 2D recipe. In 2D medical MLLMs, training often consists of aligning a mature image encoder to an LLM and then performing instruction tuning. In volumetric radiology, the central difficulty is different: the model must learn to preserve a complete volumetric study, compress it into a limited token budget, and still generate clinically grounded language from weak report-level supervision. As a result, the practical training pipeline becomes a curriculum over preservation of clinically relevant 3D information. The visual stream is first initialized with volumetric priors through SSL or 3D vision–language pre-training, the interface then learns how to expose visual tokens to the LLM, and instruction tuning finally teaches the model which parts of the scan should be verbalized, localized, or ignored for a given clinical question. ClinFusion [113] makes this curriculum explicit through five stages that progress from shallow 2D multi-encoder alignment to deep visual–language alignment, 2D instruction tuning, rapid 3D alignment, and joint 2D/3D instruction tuning.

**Volumetric initialization before instruction tuning.** Most native 3D systems therefore begin with a stronger visual initialization than their 2D counterparts. The reason is not only data efficiency, but also stability: directly coupling raw CT, MRI, or PET volumes to an autoregressive decoder would force the model to learn anatomy, spatial continuity, modality appearance, token compression, and clinical language simultaneously. E3D-GPT [120] makes this dependence explicit by first learning a self-supervised 3D foundation model before instruction tuning. M$^3$D-LaMed [117] and 3D-CT-GPT [119] rely on 3D vision-language or CT-CLIP-style pre-training so that the visual features already carry report-level semantics. Merlin [3], CT-CHAT [19], and Med3DVLM [30] reflect the same pattern: the model is not trained to discover radiological perception from SFT alone; it imports a volumetric or semantically aligned encoder and then trains the vision-language interfaces around it. MedReCo-VLM [114] provides a comparison-oriented variant: its modality-aware encoder is first trained with image–report alignment and entity-conditioned contrastive ranking, then frozen while a projector and Qwen2.5 decoder learn comparative generation from paired studies. This makes volumetric MLLM training a continuation of the representation-learning and vision–language alignment stages described in Sections 3.1 and 3.2 rather than an isolated final step.

**Volumetric instruction data construction.** The second difference is the form of supervision. A 2D medical VQA example can often be treated as an image-question-answer triple, but a volumetric radiology case is usually paired with a study-level report that does not explicitly label every slice, organ, lesion, or negative finding. Training data must therefore be converted into instructions that preserve the structure of the scan. M$^3$D-LaMed [117] scales 3D instruction-response data across multiple tasks, while CTInstruct

Table 3: Comparison of generative backbone paradigms for volumetric radiology MLLMs. The rows summarize the capability inherited by each paradigm, its dependence on the encoder-interface pipeline, and the corresponding evaluation priority.

| Generative backbone paradigm | Capability inherited | Volumetric dependency or limitation | Key evaluation implication |
|---|---|---|---|
| **Report-generation or unified token decoders** | Task-specific structured report generation or joint visual–text token modeling | Task-specific output scope; grounding depends on the encoded visual states delivered directly to the decoder | Assess report completeness, factuality, and spatial fidelity for the intended task |
| **Text-decoder-centric MLLMs** | Language fluency, instruction following, and medical terminology | Volumetric grounding remains external to the decoder and depends on encoder-interface delivery | Assess grounded factuality and hallucination jointly with encoder-interface quality rather than parameter scale alone |
| **Pretrained VLM/MLLM foundation stacks** | Multimodal conditioning, visual-token interfaces, and broader context handling | Inherited token-format and positional assumptions may be optimized for 2D images or video | Verify 3D token compatibility, spatial provenance, and grounding after volumetric adaptation |

[110] organizes CT supervision around diagnosis, segmentation, report generation, and multiple-choice reasoning. CT2Rep [115], BrainGPT [118], MEPNet [123], Brain3D [129], and PETRG-3D [128] show a more report-centric version of this strategy, where the model learns to translate full-volume information into structured radiological language. Accordingly, instruction construction in 3D is not merely prompt formatting; it determines whether the supervision teaches global impressions only, or also laterality, anatomical location, lesion extent, modality-specific cues, and uncertainty.

**Grounding-aware multi-task training.** Because study-level reports can be too coarse, many volumetric MLLMs introduce auxiliary supervision that forces generated language to remain connected to spatial grounding targets. This creates a training pattern that is less common in general 2D MLLMs: report generation is combined with classification, localization, segmentation, region description, or mask-conditioned reasoning. Reg2RG [121] uses region and mask information so report text can be tied to local image-grounded information. MedRegion-CT [125] uses pseudo-mask-derived region tokens to reduce region-level hallucination. PETAR [127] trains with PET, CT, and lesion-mask information so the response focuses on the highlighted metabolic abnormality rather than the whole scan generically. U-VLM [224] progressively supervises segmentation, classification, and report generation, while MedVL-SAM2 [221] connects language reasoning to prompt-driven 3D segmentation. UniReason-Med [218] further frames grounded reasoning as a shared interface across 2D images and slice-serialized 3D volumes. RadSight [112] organizes this supervision into a four-stage evidence hierarchy, visual–language alignment, fine-grained attribute and spatial perception, clinical diagnosis, and report generation, so diagnostic interpretation is introduced only after lesion attributes and 2D/3D grounding have been learned explicitly. CLarGen [241] identifies template collapse as a failure mode of 3D CT report generation and separates clinical detection from language synthesis. These methods suggest that, for volumetric radiology, SFT should include grounding supervision rather than only response imitation: the model must learn not just the surface form of a plausible report, but where each statement comes from in the volume.

**Token-budget-aware training.** A third distinction is that training must account for the cost of visual tokens. A 3D scan contains far more tokens than an LLM can consume, so token selection is part of the training task rather than a pure inference optimization. Hybrid slice-sequence systems such as Hulu-Med [28], Fleming-VL [227], and MedGemma 1.5 [31] train on long image sequences or video-like representations to make 2D backbones usable for volumetric studies. Efficiency-oriented models make this more explicit: MedPruner [228] learns hierarchical slice and patch pruning, Photon [108] learns instruction-conditioned visual token scheduling, and BTB3D [225] redesigns volumetric tokenization so compact tokens can be learned before LLM injection. Med-2E3 [107] and mpLLM [126] further show that token routing can be question- or modality-dependent. Thus, the 3D training objective asks a more constrained question than 2D alignment: which subset of visual tokens should survive for this instruction, and how should the model be penalized when the discarded information is critical for answering the clinical question?

**Parameter-efficient and reward-driven post-training.** Finally, 3D medical instruction data remains limited and expensive for unrestricted end-to-end tuning in most settings. Many systems therefore freeze the vision encoder or LLM, train only the vision-language interfaces or lightweight adaptation modules, and add LoRA or other parameter-efficient components to control overfitting and preserve the language model's clinical fluency. This is especially important for domain-specialized settings such as brain MRI, multiparametric MRI, and PET, where the model must learn modality-specific information without degrading general instruction-following ability. Reward-driven post-training is still emerging, but it is conceptually important for 3D because some targets are more verifiable than open-ended dialogue. Med3D-R1 [29] extends SFT with RL-style optimization for volumetric abnormality diagnosis, suggesting that rewards can be defined around diagnostic correctness, spatial grounding, response format, and clinical consistency. Recent reward designs make this visual constraint more explicit. TIF-GRPO [242] regulates anatomy-aware rewards for volumetric CT analysis, while E-MRL [132] trains diagnosis-localization-verification trajectories with cross-view consistency rewards. The value of RL-style post-training in volumetric radiology is therefore not simply to produce longer reasoning chains, but to penalize fluent statements that are unsupported by the volume and reward answers that remain anatomically and diagnostically grounded.

Overall, the distinguishing feature of volumetric MLLM training is not the mere presence of pre-training, SFT, or RL, but the way these stages are reorganized around volumetric information preservation and spatial grounding. A model must first acquire 3D anatomical and semantic priors, learn a compressed visual–language interface, and develop the ability to answer with an appropriate level of spatial detail before being optimized for correctness and grounding under clinical constraints. This makes training strategy inseparable from the encoder and interface choices discussed above: in volumetric radiology, what the model learns is tightly constrained by the slices, regions, modalities, and masks that the training pipeline teaches it to preserve.

In summary, the model landscape for volumetric radiology is moving along two coupled axes. Representation learning is evolving from generic reconstruction and contrastive learning toward anatomy-aware, fine-grained, and knowledge-enhanced alignment. MLLM construction is evolving from simple 2D encoder–LLM coupling toward native or hybrid 3D tokenization, richer vision-language interfaces, stronger generative backbones, and grounding-aware post-training. ClinFusion's retrieval and specialist-tool extension [113] illustrates the resulting bridge to agentic systems: Once volumetric information can be represented, compressed into visual tokens, spatially grounded, and discussed in language, the next system-level challenge is to orchestrate these capabilities into reliable multi-step clinical workflows.

# 4 Agentic Systems in Volumetric Radiology

The volumetric radiology MLLMs reviewed in Section 3.3 enable language-mediated interpretation of volumetric imaging, but many remain single-pass input–output models: they encode a study, expose a compressed set of visual tokens to a generative backbone, and produce an output from fixed context. Agentic systems extend this design by embedding an LLM, MLLM, or VLM within a structured workflow that supports iterative evidence acquisition, planning, tool use, and context management. This relation is better understood as co-evolution rather than replacement: stronger volumetric representations expand the evidence available to agentic systems, whereas agentic workflows impose additional requirements for representation fidelity, provenance, and traceability. Section 3 focused on how volumetric representations are learned, aligned, and preserved; this section examines how they are dynamically accessed, supplemented, and coordinated across multiple stages of analysis.

Architecturally, these systems pursue multimodal reasoning through two complementary paths. Some augment MLLM perception directly with planning, memory, and iterative refinement. Others route perceptual tasks using the language model as an indirect multimodal controller that acquires volumetric information without an internal visual encoder. Both paths extend the multimodal reasoning paradigm of Section 3 to workflow-level systems, and both share the same requirement: clinically relevant information must remain spatially grounded and traceable to their source, whether it originates from an internal encoder or an external tool call.

These characteristics motivate a functional taxonomy centered on the operations required for volumetric radiology analysis. Because clinically relevant evidence may be distributed across slices, series, phases, anatomical regions, measurements, and prior examinations, an agent must solve four linked problems: deciding what evidence to inspect or acquire; translating clinical intent into spatial or quantitative operations; retaining and updating evidence across analysis steps and time points; and coordinating intermediate outputs with tools, other agents, and human review. These problems correspond, respectively, to reasoning and planning, tool-augmented perception and grounded action, memory and dynamic context management, and workflow interaction and multi-agent collaboration.

We use these four capability families as a functional taxonomy and examine each through a common analytical lens: its role in volumetric radiology analysis, the design patterns used to implement it, and the scope and strength of its validation. Figure 6 illustrates how these capabilities form an iterative analysis loop, while Table 4 maps representative systems to their dominant functional roles. The following subsections examine their design patterns and validation boundaries in greater detail.

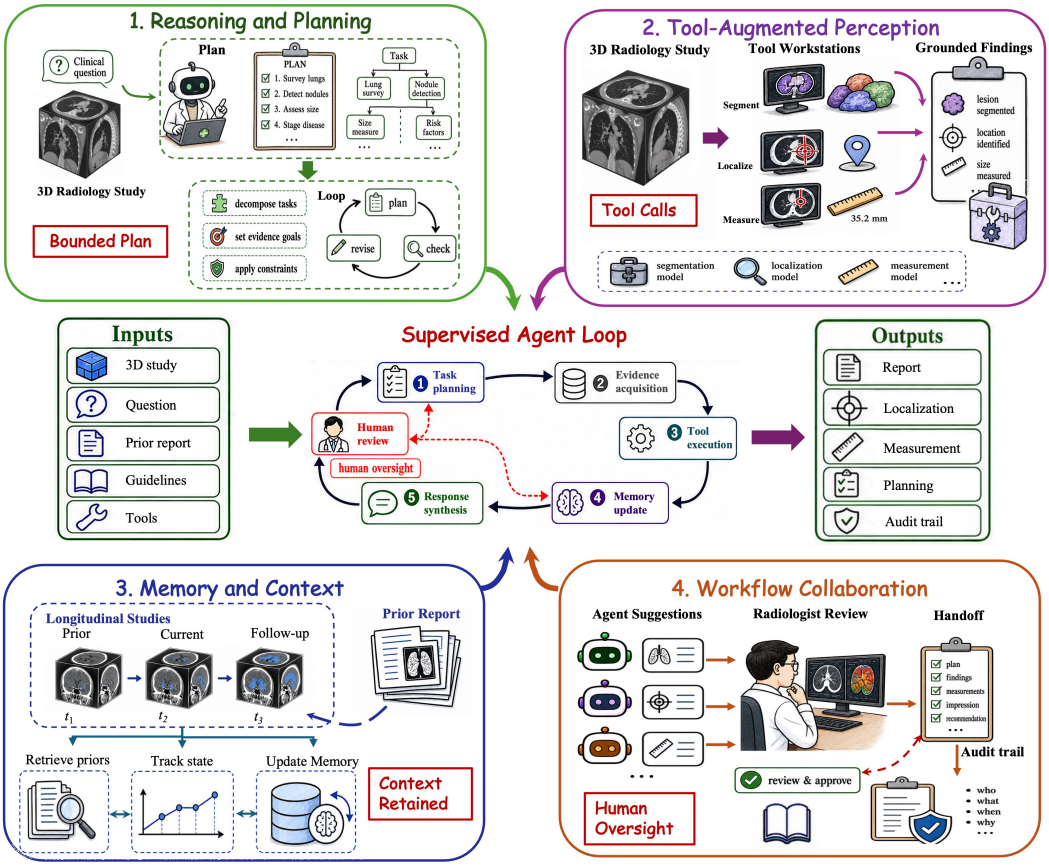


Figure 6: **Agentic workflow and modules for volumetric radiology analysis.** The four capability families correspond to the main operations required for volumetric analysis: selecting and sequencing evidence, converting clinical intent into spatial or quantitative actions, retaining context across analysis steps and examinations, and coordinating intermediate outputs within supervised workflows. Solid arrows denote workflow progression, dashed arrows denote feedback, and red tags indicate supervision or operational constraints.

## 4.1 Reasoning and Planning

This subsection examines reasoning and planning as the control layer for navigating distributed volumetric evidence. It considers how agents select the anatomy, slices, series, or phases to inspect; determine whether additional evidence or a tool call is required; and synthesize local observations into a study-level conclusion. The discussion compares anatomy-aware decomposition, procedure-guided interpretation, policy-constrained evidence acquisition, and planner–executor control, and evaluates how the resulting plans and tool-use trajectories are validated.

**Capability definition and radiological role.** Reasoning and planning refer to the ability of an agent to decompose a radiological request into intermediate operations before producing an answer, report, or workflow action. In volumetric radiology, the central difficulty is that the relevant information may be distributed across slices, series, anatomical regions, measurements, and prior observations. Planning therefore acts as a control layer between volumetric information and language generation. It determines which anatomy should be inspected, which slices or views should be selected, whether a tool call is required, and how local observations should be synthesized into a study-level conclusion.

**Representative design patterns.** Four design patterns help organize the current literature. Anatomy-aware decomposition narrows a study to clinically relevant regions or slices before synthesis; CT-Agent [35] and 3DMedAgent [36] illustrate this route through global-to-local CT analysis, informative-slice selection, and aggregation of intermediate observations. Procedure-guided interpretation makes the reading process explicit. RadAgent [37] uses stepwise chest CT understanding, while MedScribe [133] starts from candidate hypotheses and acquires localized observations through pathology-specific tools before

Table 4: Representative agentic systems for volumetric radiology analysis, organized by the primary capability emphasized by each system. MCP: Model Context Protocol

| Representative systems | Main capability | 3D-specific design | Input and output | Evaluation focus | Main limitations |
|---|---|---|---|---|---|
| *Reasoning and planning* | | | | | |
| CT-Agent [35], 3DMedAgent [36], RadAgent [37], MedScribe [133], MedOpenClaw [134], MedVistaGym [135], PD-CTAgent [136] | Decompose volumetric analysis into intermediate reasoning, targeted inspection or evidence acquisition, tool use, and synthesis steps | Global-to-local CT analysis, phase-aware evidence sufficiency, iterative phase escalation, full-study navigation, checklist-guided interpretation, and tool-integrated reasoning | Full radiology studies or partial multi-phase evidence, with outputs including answers, phase requests, localized findings, reports, and reasoning or tool-use trajectories | VQA accuracy, report quality, diagnostic and phase-escalation accuracy, reasoning trace, tool-use trajectory, expert review | Reasoning traces and acquisition policies are mostly evaluated indirectly or retrospectively and are not yet reliable indicators of faithful clinical control |
| *Tool-augmented perception and grounded action* | | | | | |
| MedToolica [137], Neuro-Radiological Agent [138], BAAI Cardiac Agent [139], CTPA-Agent [140], GAZE [141], AgentMRI [142], NeuroAgent [143], NEXUS [144], VoxelPrompt [145], MedSAM-Agent [146], MedSegAgent [147] | Invoke specialized perception, reconstruction, measurement, retrieval, viewer, or pipeline-orchestration tools | Quantitative abdominal CT reasoning, 3D masks, volumetry, cardiac indices, CTPA abnormality labels, MRI correction, viewer operations, multimodal preprocessing pipelines, multi-agent workflow orchestration, and promptable segmentation | Volume or imaging study to mask, measurement, corrected image, abnormality label, answer, or report | Quantitative reasoning, segmentation, detection, AUROC, reconstruction quality, grounding quality, tool-call validity | Final performance depends on tool reliability, tool selection, spatial parameterization, and failure detection |
| *Memory and dynamic context management* | | | | | |
| 3DMedAgent [36], Neuro-Oracle [148], Agent-MIRA [149], TheraAgent [150], BT-RADS Agent [151], LungNoduleAgent [152] | Retain intermediate observations, retrieved cases, prior studies, longitudinal lesion state, or treatment context | Similar-case PET retrieval, pre-to-postoperative MRI trajectories, longitudinal tumor response, lesion follow-up, and shared multi-agent memory | Current and prior imaging with reports or clinical text to diagnosis, prognosis, response category, or report | Prognostic accuracy, response agreement, score agreement, retrieval relevance, memory ablation | Memory quality, source attribution, temporal comparability, and lesion correspondence remain insufficiently validated |
| *Workflow interaction and multi-agent collaboration* | | | | | |
| Radiologist Copilot [39], MARCH [153], CT-Flow [38], SpineAgent [154], DosimeTron [155], GPT-Plan [156], DOLA [157], MARTP [158], SAGE [159], Scan-do Attitude [160], PET/CT Agent [161], NEXUS [144] | Coordinate reporting, protocol management, dosimetry, treatment planning, neuroimaging research execution, or end-to-end imaging workflows | Role-specialized agents, MCP-based orchestration, multi-sequence MRI reporting, DICOM selection, PET/CT quantification, dose optimization, composable primitives, and constraint checking | Clinical workflow inputs to draft report, protocol action, dose estimate, treatment plan, or decision-support output | Workflow completion, expert preference, dose or plan quality, processing time, correction burden | Prospective validation, workload measurement, failure recovery, and human handoff are still incompletely studied |

report synthesis. Policy-constrained evidence acquisition treats planning as a decision about whether the current evidence is sufficient. PD-CTAgent [136] abstracts each available CT phase into a structured evidence package and combines guideline retrieval, uncertainty signals, and rule-prioritized VLM control to request another phase only when the current state is judged insufficient. Planner-executor control appears in systems that link imaging findings to management, guideline, or interactive-viewer actions, including INFORM-CT [185], ACR-oriented reasoning agents [243], MedOpenClaw [134], and MedVistaGym [135]. These patterns differ less in their nominal reasoning label than in the granularity of the exposed plan: region selection, clinical checklist, evidence-acquisition policy, management action, or interactive tool trajectory.

**Reported results and practical constraints.** Reported results support planning mainly as a practical control mechanism. Current systems are usually evaluated through final VQA accuracy, report quality, expert review, task completion, or logged tool trajectories. These endpoints show whether a planned workflow can improve observable performance or make the process more inspectable. They do not establish that the final statement is causally supported by the intermediate steps. A plausible trace may contain unnecessary actions, weakly grounded observations, or retrospectively convincing language. The current conclusion should therefore remain narrow: planning helps organize targeted inspection and tool use, but trace fidelity and clinical reasoning validity remain incompletely tested.

## 4.2 Tool-Augmented Perception and Grounded Action

This subsection examines tool-augmented perception as the mechanism that converts language-mediated intent into explicit spatial or quantitative operations on a volumetric study. It distinguishes specialist perception tools, viewer or retrieval tools, and language-to-spatial-action interfaces, and then considers tool governance and failure detection.

**Capability definition and radiological role.** Tool-augmented perception refers to the use of external modules for operations that language generation cannot perform reliably by itself. In volumetric radiology, these operations include segmentation, registration, reconstruction correction, measurement, viewer manipulation, retrieval, DICOM handling, standardized uptake value (SUV) conversion, and dose simulation. The role of tool use is to turn a textual or clinical request into explicit spatial grounding targets or quantitative outputs, such as a mask, measurement, corrected image, selected series, or dose map. A statement about a pulmonary embolus, enhancing lesion, ventricular volume, or absorbed dose is clinically interpretable only when it is tied to the correct image region, coordinate frame, series, or measurement pipeline.

**Representative design patterns.** Tool-augmented perception supports validation when the tool output becomes part of a traceable path from image data to final claim. A first pattern is specialist tool coordination, in which an agent delegates spatial or quantitative operations to domain-specific modules. MedToolica [137] applies this principle to quantitative 3D abdominal CT reasoning, coordinating expert tools for organ-centric measurement, pathology assessment, spatial comparison, and clinical interpretation. PD-CTAgent [136] uses organ-aware segmentation and disease-specific tool routing to construct evidence packages that include lesion measurements, quality-control flags, and structure-aware uncertainty before phase-sufficiency reasoning. The Neuro-Radiological Agent [138] combines skull stripping, registration, tumor segmentation, volumetry, and response assessment for brain MRI, and the BAAI Cardiac Agent [139] integrates cardiac MRI segmentation, functional quantification, tissue characterization, diagnosis, and reporting. In nuclear medicine, DosimeTron [155] automates DICOM metadata extraction, PET/CT preprocessing, Monte Carlo simulation, organ segmentation, and dosimetric reporting, whereas the PET/CT Agent [161] coordinates raw DICOM series selection, registration and resampling, SUV conversion, segmentation, maximum-intensity projection or fusion-image generation, vision-enabled understanding, and structured draft reporting. A second pattern gives the agent access to viewer, reconstruction, retrieval, or preprocessing tools. AgentMRI [142], GAZE [141], and NeuroAgent [143] illustrate this broader information-access role. A third pattern treats language as an interface for spatial action, as in VoxelPrompt [145], MedSAM-Agent [146], MedSegAgent [147], MedSAM3 [244], and IBISAgent [245]. The key distinction among these patterns is the form of checkable output: numeric measurements and masks are more directly verifiable than viewer manipulations or retrieved contextual material.

**Reported results and practical constraints.** Support for tool-augmented perception is most direct when the tool produces a directly measurable output. Masks, volumetry, cardiac indices, reconstruction corrections, abnormality labels, and dose estimates can be assessed with task-specific metrics. The central unresolved problem is governance of the tool chain. An agent may choose the wrong tool, pass incorrect parameters, accept a flawed segmentation, use the wrong DICOM series, or combine incompatible outputs. ToolSelect [246] and tool-expertise-aware agents [247] highlight this issue in broader medical-agent settings. In volumetric radiology, such errors are amplified by spatial dependence: an inaccurate registration can change longitudinal response assessment, and an incorrect mask can alter dose or volume estimates. Tool validity therefore has to be evaluated both at the module level and at the workflow level, where tool selection, parameterization, failure detection, and final claim attribution are all visible.

## 4.3 Memory and Dynamic Context Management

This subsection examines memory and dynamic context management as the mechanism for integrating evidence across slices, series, analysis steps, retrieved cases, and longitudinal examinations. It distinguishes working memory, longitudinal patient memory, and retrieval-augmented memory, and then considers provenance, comparability, and lesion correspondence.

**Capability definition and radiological role.** Memory and context management refer to the ability of an agent to retain, retrieve, and update information that is not contained in a single model input. In volumetric radiology, this capability is required because a study can include many slices and series, and a patient may also have prior scans, prior reports, treatment history, and longitudinal lesion measurements. Memory supports intra-study aggregation, prior-study comparison, similar-case retrieval, and response assessment

over time. It also addresses a practical limitation of direct MLLM inference: not all visual tokens, reports, measurements, retrieved cases, and tool outputs can remain in the active context simultaneously.

**Representative design patterns.** Current memory designs differ by the time scale and source of the stored information. Working memory operates within a single analysis episode. CT-Agent [35] and 3DMedAgent [36] store selected regions, intermediate observations, retrieved context, or tool outputs while analyzing a study. PD-CTAgent [136] records a compact phase-by-phase state containing the active CT phase, structured evidence package, policy decision, and matched guideline trace, enabling newly requested phases to be reprocessed rather than merely appended to the language context. Longitudinal patient memory is more clinically distinctive. Neuro-Oracle [148] models pre-to-postoperative MRI trajectories for seizure prognosis, TheraAgent [150] uses self-evolving memory for PET theranostic response prediction, BT-RADS Agent [151] supports brain-tumor follow-up scoring, and LungNoduleAgent [152] uses shared memory and multi-agent discussion for lung nodule diagnosis. Retrieval-augmented memory expands the context beyond the current case, as in Agent-MIRA [149], GAZE [141], and ACR guideline agents [248]. These patterns are not interchangeable: working memory supports local aggregation, longitudinal memory supports change assessment, and retrieval-augmented memory supports comparison with external cases or guidance.

**Reported results and practical constraints.** Memory-related claims are usually supported by task-level outcomes such as prognostic accuracy, response agreement, score agreement, retrieval relevance, ablation of memory components, or longitudinal consistency. These signals indicate whether stored context helps the target task, but they rarely isolate the memory module from the perception model, retrieval system, or decision layer. Stored or retrieved information may also be inaccurate, outdated, or non-comparable. A prior report may contain an error, a retrieved PET case may differ in disease stage, and lesion correspondence may fail after surgery or treatment. For retrieval systems, nearest-neighbor similarity in feature space does not necessarily imply clinical comparability. For longitudinal systems, valid temporal reasoning depends on registration quality, lesion matching, and treatment timing. Memory therefore strengthens agentic radiology only when source attribution, update rules, and comparability are visible to the reviewer.

## 4.4 Workflow Interaction and Multi-Agent Collaboration

This subsection examines workflow interaction and multi-agent collaboration as the coordination layer linking models, tools, intermediate artifacts, and human review across volumetric radiology procedures. It covers reporting, planning and operational workflows, reproducible multi-agent execution, supervised assistance, handoff, and workload validation.

**Capability definition and radiological role.** Workflow interaction and multi-agent collaboration refer to systems that place LLMs, MLLMs, or VLMs within clinical, technical, or research procedures. Instead of producing a single answer, the system coordinates roles, tools, intermediate artifacts, and human review. This capability is relevant to volumetric radiology because reporting, dosimetry, treatment planning, protocol management, spine MRI understanding, and neuroimaging analysis all depend on multi-step procedures with handoffs between people, software, and quantitative outputs. In this setting, the agent functions as an organizer of the workflow rather than only an image interpreter.

**Representative design patterns.** Workflow-oriented agents differ from task agents because the unit of operation is a clinical or technical process rather than a single prediction. Reporting workflow assistance is the first pattern. Radiologist Copilot [39] coordinates reporting tools with quality control, MARCH [153] models a resident-fellow-attending hierarchy for CT report generation, and CT-Flow [38] uses model context protocol servers to make interpretation state and tool invocation explicit. SpineAgent [154] extends this reporting logic to multi-sequence spine MRI by combining a multi-sequence foundation model with specialized agents for diagnosis, pathological-region localization, similar-case retrieval, and a medical report agent. A second pattern is planning and operational workflow assistance. PD-CTAgent [136] conditions evidence-sufficiency decisions and report emphasis on an active institutional rule set; its scope is to decide whether additional phase evidence is needed from partially observed studies, rather than to prescribe acquisition autonomously. DosimeTron [155], GPT-Plan [156], DOLA [157], MARTP [158],

and SAGE [159] show how agentic control can assist PET/CT dosimetry or radiotherapy planning, where dose estimates, constraints, and role-specific checks matter more than open-ended dialogue. A third pattern focuses on reproducible execution, including NeuroClaw [249], artifact-based medical image processing agents [250], and NEXUS [144]. These examples show that agentic workflows can support upstream acquisition or reconstruction tasks, downstream interpretation and reporting, and research-oriented processing pipelines.

**Reported results and practical constraints.** Workflow agents are supported by outcomes matched to their intended use: report quality, expert preference, RADPEER-style review, dose agreement, processing time, plan quality, constraint satisfaction, or reproducible execution logs. SpineAgent illustrates this standard by combining automated report-generation metrics with review by five radiologists, but even this framing remains explicitly assistive and in-scope. Across the broader literature, relatively few studies measure whether the agent selected the correct tool, used correct parameters, detected failures, preserved intermediate observations, or reduced clinician workload under realistic constraints. The relevant validation standard also changes by workflow: reporting assistants require claim accuracy and review efficiency, dosimetry agents require quantitative dose agreement and segmentation reliability, planning agents require constraint satisfaction and deliverability, and research agents require reproducibility. Current systems should therefore be interpreted as bounded and supervised assistance systems. Their near-term value is to organize complex volumetric radiology workflows and make intermediate steps more traceable, while clinical authority remains with human experts.

Together, these four capability families form an iterative volumetric-analysis loop: reasoning and planning select and sequence the evidence to inspect; tool-augmented perception produces grounded spatial or quantitative outputs; memory integrates evidence across analysis steps and examinations; and workflow interaction coordinates these outputs with clinical procedures and human review. Agentic volumetric radiology systems thereby extend or complement volumetric radiology MLLMs at the workflow level.

# 5 Clinical Applications and Evaluation in Volumetric Radiology

Clinical evaluation begins with the intended claim. The capabilities reviewed in Sections 3 and 4 do not by themselves establish clinical validity: the appropriate evaluation depends on what a system is intended to predict, communicate, or support in practice. Accordingly, validation should be matched to both the type of clinical output and the strength of the claim being made.

We therefore organize clinical applications and evaluation according to four claim-driven families: diagnostic interpretation and reporting, prognosis and treatment response, image-derived planning and clinical decision support, and cross-task evaluation and validation. These families differ in the clinical question being addressed and, consequently, in the validation required to support those claims. Within each family, we examine the intended clinical role, the corresponding evaluation strategy, and the principal limitations of current validation practice.

The strength of a clinical claim should also be commensurate with the level of validation. Internal benchmarks primarily establish technical feasibility, external or multi-institutional studies assess robustness, reader studies and workflow simulations examine clinical usability, and prospective evaluation is required to support deployment-oriented claims. This progression operationalizes the validation component of the Claim–Design–Validation framework introduced in Section 1. Figure 7 summarizes the relationship among clinical claims, evaluation strategies, and validation levels, while Tables 5 and 6 summarize representative systems and benchmark resources.

## 5.1 Diagnostic Interpretation and Reporting

**Task definition and clinical role.** Diagnostic interpretation and reporting assess whether a system can identify, localize, characterize, and communicate clinically relevant findings from a volumetric radiology study. Outputs may include spatial predictions (e.g., masks and detections), diagnostic predictions (e.g., labels or scores), and language-mediated interpretations (e.g., VQA responses, grounded findings, and structured reports). This application family supports routine interpretation, triage, lesion measurement,

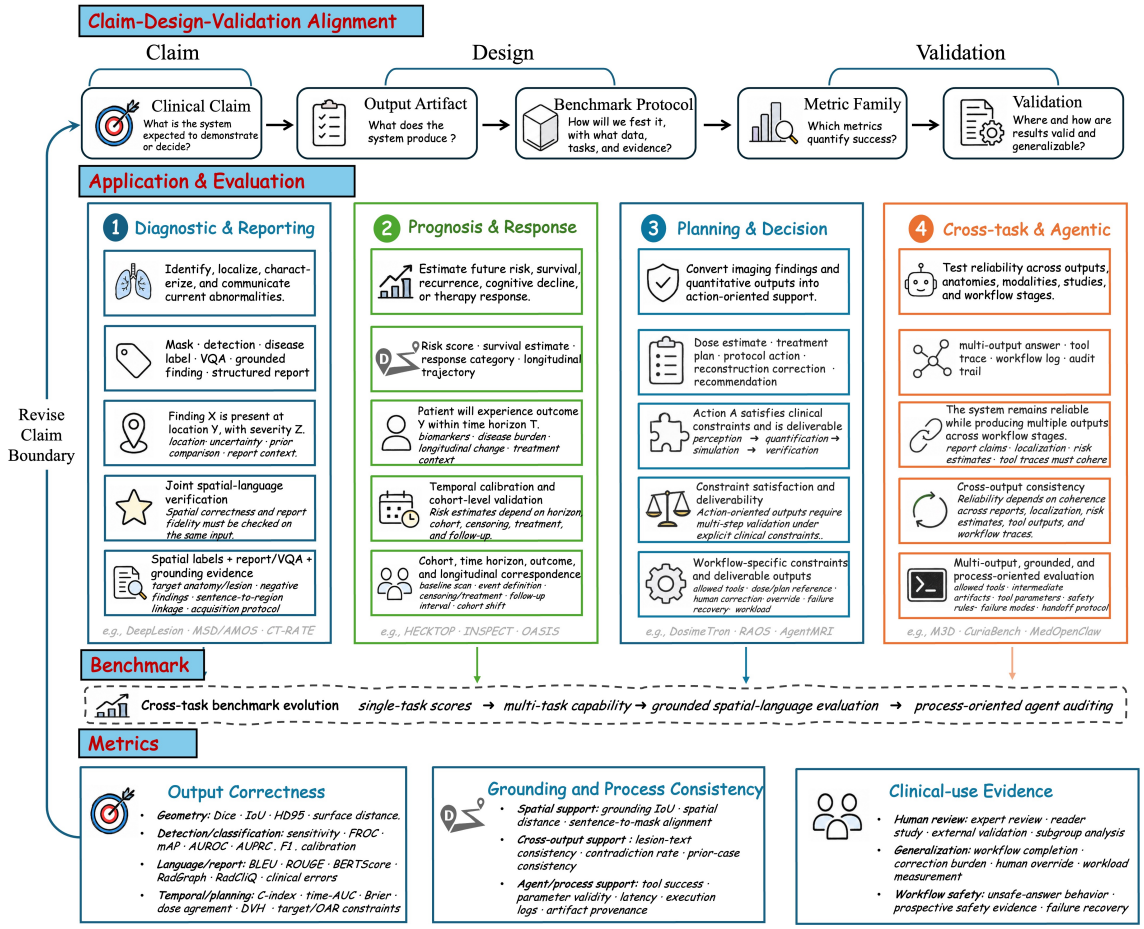


Figure 7: **Clinical applications and evaluation for volumetric radiology.** The figure maps four application families to their claim types, validation demands, benchmark protocols, and metric evidence, emphasizing that evaluation standards should follow the clinical claim and intended output.

follow-up documentation, and communication with downstream care teams. Clinical usefulness therefore requires both evidence preservation and appropriate communication: the system must retain the volumetric information supporting each claim, and the output must convey relevant attributes such as anatomical location, laterality, extent or severity, uncertainty, negative findings, and comparison with prior studies when applicable.

**Representative methods and validation demands.** Early evaluation in this family focused primarily on segmentation and detection, for which reference masks, bounding boxes, and lesion annotations made spatial correctness directly measurable. Representative datasets include LIDC-IDRI [251], LUNA16 [252], BraTS [183], MSD [165], AMOS [164], AbdomenAtlas [166], and TotalSegmentator [253]. These tasks support geometric and lesion-level evaluation using metrics such as Dice, intersection over union (IoU), the 95th percentile Hausdorff distance (HD95), free-response receiver operating characteristic (FROC) analysis, and lesion-level recall.

As 3D foundation models and MLLMs extended the output space toward diagnostic prediction, VQA, and report generation, evaluation also had to address semantic correctness, report completeness, and correspondence between language and image. Representative resources and systems include CT-RATE [19], CT2Rep [115], M$^3$D [117], Merlin [3], MedVista3D [89], 3D-BrainCT [118], and Brain3D [129]. Agentic radiology systems, including CT-Agent [35], PD-CTAgent [136], RadAgent [37], MedScribe [133], Radiologist Copilot [39], MARCH [153], and the PET/CT Agent [161], further broaden the evaluated process to include full-study navigation, targeted inspection, phase-sufficiency decisions, tool-mediated evidence acquisition, and report verification.

The required validation evidence depends on the output type. Spatial outputs require geometric

Table 5: Clinical application families and validation requirements in volumetric radiology AI. Representative systems and evaluation resources are summarized by clinical claim, task scope, commonly reported evidence, and principal validation gap. Dataset-level details are provided separately in Table 6. AUC, area under the curve; AUPRC, area under the precision–recall curve; AUROC, area under the receiver operating characteristic curve; C-index, concordance index; DVH, dose–volume histogram; FROC, free-response receiver operating characteristic; HD95, 95th percentile Hausdorff distance; IoU, intersection over union; PE, pulmonary embolism.

| **Systems/resources** | **Clinical claim** | **Representative tasks/outputs** | **Common evaluation evidence** | **Principal validation gap** |
|---|---|---|---|---|
| *Diagnostic interpretation and reporting* | | | | |
| CT-RATE [19], CT2Rep [115], $M^3D$ [117], Merlin [3], MedVista3D [89], CT-Agent [35], and MedScribe [133] | Identify, localize, characterize, and communicate clinically relevant findings | Segmentation, detection, classification, VQA, grounded findings, and report generation | Dice, IoU, HD95, and FROC; AUROC and calibration; BLEU/ROUGE, RadGraph, RadCliQ, and grounding IoU | Clinical correctness and traceability to volumetric evidence remain incompletely assessed |
| *Prognosis and treatment response* | | | | |
| CLIP-Lung [95], LoV3D [179], CTPA-Agent [140], Neuro-Oracle [148], TheraAgent [150], and BT-RADS Agent [151] | Predict future risk or clinically meaningful change over time | Malignancy and survival prediction, longitudinal response assessment, and cognitive outcomes | AUROC, AUPRC, C-index, time-dependent AUC, Brier score, calibration, and response agreement | Temporal validity, including calibration and lesion correspondence, remains insufficiently established |
| *Image-derived planning and clinical decision support* | | | | |
| DosimeTron [155], GPT-Plan [156], SAGE [159], AgentMRI [142], Scan-do Attitude [160], and the PET/CT Agent [161] | Translate imaging evidence into technically valid planning or management outputs | Dosimetry, radiotherapy planning, reconstruction and protocol control, and incidental finding management | Dose agreement, dose–volume constraints, reconstruction quality, workflow completion, and expert correction | End-to-end error propagation and recovery remain poorly characterized |
| *Cross-task evaluation and validation* | | | | |
| $M^3D$ [117], CuriaBench [200], CT-SpatialVQA [190], Med-StepBench [191], CORTEX [192], ABRA [195], MedCTA [196], and RadSaFE-200 [198] | Assess coherence across outputs and workflow stages | Multi-task VQA and reporting, spatial reasoning, grounding, tool-use evaluation, and safety assessment | Task-specific scores, grounding IoU, spatial error, QAScore, trace fidelity, workflow completion, and unsafe-answer rate | Integrated case-level validation and detection of propagated errors remain limited |

*Abbreviations:* AUC, area under the curve; AUPRC, area under the precision–recall curve; AUROC, area under the receiver operating characteristic curve; BLEU, Bilingual Evaluation Understudy; C-index, concordance index; FROC, free-response receiver operating characteristic; HD95, 95th percentile Hausdorff distance; IoU, intersection over union; ROUGE, Recall-Oriented Understudy for Gisting Evaluation; VQA, visual question answering.

and lesion-level assessment, whereas diagnostic labels require discrimination, calibration, and error analysis. Language outputs are commonly evaluated with BLEU, ROUGE, METEOR, CIDEr, BERTScore, RadGraph, RadCliQ, clinical error measures, and question-answering metrics. These metrics capture complementary dimensions of report quality, but they do not by themselves establish whether generated statements are supported by the underlying volume. High textual similarity can coexist with clinically incorrect or ungrounded claims, while spatially accurate predictions can still be communicated inaccurately. Diagnostic evaluation should therefore determine whether the resulting interpretation is clinically correct and traceable to the relevant volumetric evidence. Expert review remains necessary when these properties cannot be reliably established by automated metrics.

**Current resources and remaining gaps.** Current resources increasingly evaluate whether language outputs remain grounded in volumetric evidence. RadGenome-Chest CT [169], AutoRG-Brain [170], and ReXGroundingCT [173] link free-text findings to anatomical regions or lesion masks. CT-SpatialVQA [190] evaluates semantic and spatial reasoning over 3D anatomy, whereas ReportQA [193] assesses report utility using knowledge-tree-derived question and answer pairs and QAScore. Perception-Bench [112] traces errors from lesion attributes and 2D/3D spatial grounding through diagnosis and report generation. MedReCo-DB [114] adds entity-conditioned reference retrieval and prior–current comparative reasoning

across seven modalities, and ClinFusion [113] complements these capability tests with MedIF-Bench for structured instruction following and an RoI-grounded factuality protocol for report evaluation. Oncology VQA [194] extends this evaluation to private 3D oncology cohorts with contamination-aware testing.

Collectively, these resources broaden evaluation beyond text similarity, but each addresses only part of the evidence required for a diagnostic claim. Calibration of probabilistic predictions, reader or workflow studies, and external validation across institutions and acquisition settings remain limited.

## 5.2 Prognosis and Treatment Response

**Task definition and clinical role.** Prognosis and treatment response assess whether imaging can support the prediction of future outcomes or clinically meaningful change over time. Targets include risk and time-to-event outcomes, such as malignancy, recurrence, and survival, as well as longitudinal outcomes, such as treatment response, cognitive decline, and structured follow-up scores. These tasks support risk stratification, follow-up planning, response assessment, clinical-trial enrichment, and identification of patients who require closer longitudinal review. Unlike diagnostic interpretation, the target is often not directly observable in a single study. Reliable inference must therefore relate imaging biomarkers and disease burden to temporal change, treatment exposure, and relevant clinical context.

**Representative methods and validation demands.** Radiomics first linked handcrafted imaging features to clinical outcomes through statistical models [10, 11]. Supervised 3D deep learning later learned prognostic features directly from imaging. Representative cohorts and benchmarks include HECKTOR [181], INSPECT [180], OASIS-3 [182], BraTS [183], and Gastric-X [184], covering survival, progression, cognitive outcomes, and treatment response.

Recent approaches broaden the evidence available to prediction. CLIP-Lung [95] and AutoRad-Lung [178] apply vision–language modeling to nodule malignancy assessment, whereas LoV3D [179] models change across longitudinal brain MRI. Systems such as CTPA-Agent [140], Neuro-Oracle [148], Agent-MIRA [149], TheraAgent [150], and BT-RADS Agent [151] further incorporate longitudinal or retrieved context into outcome prediction and structured response assessment.

Validation in this family must establish more than discrimination. AUROC and the C-index indicate whether a model ranks patients by risk, but they do not establish accurate absolute risk or reliable change assessment at a clinically relevant time point. The endpoint, prediction horizon, cohort construction, and treatment exposure must therefore be specified, with censoring and competing events handled appropriately. Without these elements, apparent performance may reflect endpoint ambiguity or cohort composition rather than clinically actionable prognostic information.

**Current resources and remaining gaps.** Current resources span binary-risk prediction, time-to-event modeling, and longitudinal response assessment. Binary outcomes are commonly evaluated with AUROC, AUPRC, sensitivity, specificity, and F1 score, whereas time-to-event models use the C-index, time-dependent AUC, and Brier score. Treatment-response studies additionally require agreement with task-appropriate clinical criteria, such as the RECIST [254], PERCIST [255], and RANO [256]. Interpretation of these measures depends on clearly defined endpoints, cohorts, and prediction horizons.

The principal gap is incomplete temporal validation. Current studies often report discrimination or response agreement without establishing calibration at the intended horizon or reliable lesion correspondence across serial examinations. External validation also remains limited, leaving transportability beyond the development cohort uncertain. For retrieval- or memory-augmented systems, historical or external evidence must additionally be temporally aligned and traceable to its source.

Table 6: Representative benchmark datasets and evaluation resources for volumetric radiology models. Dataset names follow the corresponding papers or project names, and resources are grouped by the application and evaluation tasks emphasized in this section. CTPA, computed tomography pulmonary angiography; CXR, chest radiography; OAR, organ at risk; PACS, picture archiving and communication system; PSMA, prostate-specific membrane antigen; RT, radiotherapy

| Dataset Name | Task | Modality | Sample Size | Highlight |
|---|---|---|---|---|
| *Diagnostic interpretation and reporting* | | | | |
| MedMNIST v2 [162] | 2D and 3D biomedical image classification | Multiple 2D and 3D modalities | 708K 2D images, 10K 3D images | Lightweight, standardized, diverse labels |
| DeepLesion [163] | Multi-class lesion detection and categorization | CT | 33,688 annotated lesion images | PACS-mined, heterogeneous lesions |
| AMOS [164] | Multi-organ abdominal segmentation | Abdominal CT and MRI | 500 CT, 100 MRI scans | Multi-center, multi-modality, 15 organs |
| MSD [165] | Segmentation across multiple anatomical tasks | 3D and 4D CT and MRI | 2,633 3D volumes | 10 tasks, generalizability benchmark |
| AbdomenAtlas [166] | Abdominal multi-organ segmentation and transfer learning | Contrast-enhanced CT | 20,460 CT volumes | 112 hospitals, large-scale annotations |
| TotalSegmentator MRI [167] | Whole-body multi-structure segmentation | Whole-body MRI | 616 MRI examinations | Sequence-independent, 50 structures |
| AMOS-MM [164] | Abdominal report generation and VQA | Multi-phase abdominal CT | 2,300 CT scans | Image-text pairs, abdominal multimodal tasks |
| CT2Rep [115] | 3D CT report generation | Chest CT and reports | 25,701 CT volumes | 3D report generation, automated reporting |
| CT-RATE [19] | CT classification and report-grounded modeling | CT and text | 25,692 scans | Public paired 3D images and reports |
| CT-3DRRG [168] | 3D CT report generation, retrieval, and grounding | Multimodal 3D CT | – | 3DRRG benchmark, visual token compression |
| RadGenome-Chest CT [169] | Grounded report generation, VQA, organ segmentation | Chest CT and aligned text | 25,692 CTs, 1.3M VQA pairs | Region-guided, 197 organ masks |
| AutoRG-Brain [170] | Brain MRI report generation and region grounding | Brain MRI and reports | 3,408 imaging-report pairs | Grounded brain MRI reporting |
| 3D-BrainCT [118] | Brain CT report generation and MLLM evaluation | Brain CT and text | 18,885 text-scan pairs | FORTE metric, brain CT reporting |
| DeepTumorVQA [171] | Tumor-centric VQA and clinical reasoning | Abdominal CT | 9,262 CT volumes | Small tumor detection, 4 QA categories |
| NOVA [172] | Anomaly localization and step-wise reasoning | Brain MRI and text | 900 scans | Rare pathologies, stress-test setting |
| ReXGroundingCT [173] | Free-text finding grounding and lesion localization | Chest CT | 3,142 CT scans | Sentence-level grounding, 3D masks |
| ViPET-ReportGen [174] | PET/CT report generation and auxiliary classification | PET/CT and text | 2,757 PET/CT volumes | Whole-body functional imaging, Vietnamese reports |
| SGMRI-VQA [175] | Detection, localization, classification, counting, diagnosis | Volumetric MRI | 41,307 QA pairs | Multi-frame reasoning, box supervision |
| SpatialMed [176] | 3D spatial reasoning and volume estimation | CT and text | 10K QA pairs | Spatial reasoning, numerical estimation |
| CT-SpatialVQA [190] | Semantic-spatial VQA | CT and text | 9,077 QA pairs from 1,601 CT volumes and reports | Anatomical localization, laterality, structural comparison, 3D relational reasoning |
| CORTEX [192] | Structured reasoning, VQA, and report generation | Chest CT and text | 76,177 validated reasoning traces | Four-stage diagnostic traces, stage-wise rubric scoring, expert review |
| PET-CLIP Captioner [177] | Whole-body PET/CT lesion captioning | Whole-body PET/CT | 1,867 subjects | Private cohort, location-guided captioning |
| Med-StepBench [191] | Step-wise hallucination detection | Oncological PET/CT | >12,000 images and >1,000,000 image-statement pairs | Four-stage diagnostic reasoning, clinician-verified hallucination evaluation |
| Oncology VQA [194] | Report-derived oncology VQA | 3D oncology CT and MRI | 2,511 cases and 42,997 questions | Private-cohort benchmark, RADS-style and report-derived questions, blind ablation |
| ReportQA [193] | QA-based report evaluation | CXR, brain CT, chest CT, abdominal CT | 6,857 reports and approximately 660,000 QA pairs | QAScore, knowledge-tree report evaluation, multi-modality report utility |
| *Prognosis and treatment response* | | | | |
| INSPECT [180] | Pulmonary embolism diagnosis and prognosis | CTPA, CXR, clinical data | 19,402 patients | Longitudinal records, multiple outcomes |
| HECKTOR [181] | Tumor segmentation and outcome prediction | Head-and-neck PET/CT | 224 cases | Survival prediction, PET/CT tumors |
| OASIS-3 [182] | Longitudinal neuroimaging and cognitive outcomes | Brain MRI, PET, clinical data | 1,098 participants | 15-year tracking, Alzheimer's cohort |
| BraTS series [183] | Brain tumor segmentation and outcome-oriented tracks | Multi-parametric brain MRI | Challenge series | Expert annotations, multi-track evaluation |
| 3D-RAD [9] | Multi-temporal VQA and longitudinal diagnosis | 3D CT | 136,195 expert-aligned samples | Static and temporal diagnosis tasks |
| Gastric-X [184] | Gastric cancer detection, staging, response assessment | Multi-phase CT, endoscopy | 1.7K cases | Multi-phase CT, multimodal cancer benchmark |

Table 6: Representative benchmark datasets and evaluation resources for volumetric radiology models. Continued.

| Dataset Name | Task | Modality | Sample Size | Highlight |
|---|---|---|---|---|
| *Image-derived planning and clinical decision support* | | | | |
| CT-FlowBench [38] | Quantitative, spatial, and diagnostic tool-use reasoning | CT | – | Agentic workflow, tool invocation |
| RAOS [187] | Robust abdominal organ segmentation for radiotherapy scenarios | Abdominal CT | 413 CT scans | Challenging cases, organ hallucination |
| HaN-Seg [188] | Head-and-neck organ-at-risk segmentation for RT planning | CT and T1-weighted MRI | 42 training, 14 test cases | MultimodalOAR segmentation |
| AAPM-RT-MAC [189] | Normal tissue auto-segmentation for MR-guided RT planning | T2-weighted head-and-neck MRI | 55 patients | Expert contours, RT grand challenge |
| DosimeTron cohort [155] | Patient-specific Monte Carlo internal dosimetry | PSMA PET/CT | 597 studies, 378 patients | System-level cohort, dosimetry validation |
| AgentMRI evaluation set [142] | MRI artifact detection and reconstruction model selection | Brain MRI | 150 subjects, 1,978 slices | Reconstruction control, degradation-specific tasks |
| Scan-do Attitude protocol set [160] | CT protocol configuration and protocol-file generation | CT protocol metadata | – | System-level set, protocol management |
| *Cross-task evaluation and validation* | | | | |
| M3D [117] | Retrieval, report generation, VQA, positioning, segmentation | CT, MRI, multimodaldata | 120K image-text pairs, 662K instructions | Generalist volumetric MLLM benchmark |
| MedVL-CT69K [81] | Zero-shot abnormality detection and report generation | CT | 272,124 CT scans | Fine-grained CT VLP, anatomy-level labels |
| Triad [199] | Segmentation, classification, registration | 3D MRI | 131,170 MRI volumes | Large-scale MRI pretraining, 25 downstream datasets |
| Merlin-Abdominal-CT [3] | Segmentation, lesion captioning, disease prediction | CT and text | 25,494 CT-report pairs | Pan-anatomy CT, long-context reasoning |
| CuriaBench [200] | Segmentation, regression, survival prediction, lesion detection | MultimodalCT and MRI | - | Foundation-model transfer, volumetric evaluation |
| ABRA [195] | Radiology-agent workstation interaction | Chest CT and breast MRI | 655 tasks | OHIF/Orthanc environment, DICOM navigation, planning-execution-outcome scoring |
| MedCTA [196] | Clinical tool-agent evaluation | Radiology images, pathology slides, reports | 107 clinical tasks | Clinician-verified tool trajectories, tool selection, argument validity, rollout reliability |
| Perception-Bench [112] | Attribute judgment, spatial grounding and understanding, diagnosis, anomaly detection, and report generation | 2D radiology images and 3D CT | 1.13M samples | Six-level evidence hierarchy linking low-level perception to clinical interpretation |
| MedReCo-DB [114] | Entity-conditioned retrieval and comparative VQA across reference cases and longitudinal studies | CXR, CT, MRI, and ultrasound | >690K images, >160K patients | Eight institutions, seven modalities, entity-aware cross-image comparison |
| MedIF-Bench [113] | Medical instruction-following and structured-output compliance | Mixed text and 2D/3D medical imaging tasks | 900 samples | Clinically relevant formatting constraints with rule-based compliance scoring |

## 5.3 Image-Derived Planning and Clinical Decision Support

**Task definition and clinical role.** Image-derived planning and clinical decision support assess whether imaging evidence can be translated into quantitative or operational outputs that inform subsequent clinical or technical decisions. These outputs include dose estimates, candidate treatment plans, protocol adjustments, and management recommendations. Their role is to provide technically grounded outputs for review by relevant clinical and technical specialists. Because such outputs may influence image acquisition, follow-up, or treatment delivery, evaluation must establish both technical validity and suitability for the intended workflow.

**Representative methods and validation demands.** Evaluation in this family is governed by technical validity, constraint satisfaction, and workflow compatibility rather than linguistic fluency. Earlier workflows relied on rules, manual contouring, optimization engines, and institutional protocols. Deep learning automated individual steps such as organ-at-risk segmentation, dose prediction, synthetic CT, and reconstruction correction, with RAOS [187], HaN-Seg [188], and AAPM-RT-MAC [189] providing benchmarks for anatomical delineation under planning constraints. Recent work shifts from component-level automation toward workflow orchestration. DosimeTron [155] automates PET/CT internal dosimetry, GPT-Plan [156], DOLA [157], MARTP [158], and SAGE [159] coordinate radiotherapy or stereotactic radiosurgery planning, AgentMRI [142] selects correction models for MRI degradation, Scan-do Attitude [160] manages CT protocols, Agent4MR [186] explores MR sequence development, INFORM-CT [185] manages incidental findings, and the PET/CT Agent [161] spans raw DICOM processing to structured staging support.

Validation must extend beyond component accuracy. A correct segmentation or dose estimate is insufficient if downstream constraints are not met or the output cannot be integrated into the intended workflow. DosimeTron [155], for example, reports multi-scanner dosimetric validation across 597 studies from 378 patients, whereas several recent planning agents have been evaluated in small retrospective cohorts [157, 159]. Variation in study scale and design limits cross-system comparison and constrains the strength of workflow-level claims.

**Current resources and remaining gaps.** Current resources span component-level benchmarks, system-level cohorts, and workflow-level assessments [38, 155, 187]. These settings evaluate different parts of the imaging and planning pipeline, but they are rarely linked within a common protocol. Consequently, strong performance on an isolated component does not establish reliable end-to-end operation.

The principal gap is the limited evaluation of error propagation and recovery. Most studies do not establish whether upstream failures are detected before they affect downstream decisions, or how much expert correction is required. External and prospective workflow studies are therefore needed before claims of clinical utility or deployment readiness can be justified.

## 5.4 Cross-Task Evaluation and Validation

**Task definition and clinical role.** Cross-task evaluation extends task-based evaluation from individual clinical outputs to systems that perform multiple tasks across anatomies, modalities, and workflow stages. Each constituent task should retain a clearly specified clinical claim, target population, imaging process, reference standard, and task-appropriate figure of merit [257, 258]. Broad task coverage alone should not be interpreted as evidence of generalist competence if performance on clinically important tasks is degraded. General-Level operationalizes task-level synergy as a generalist surpassing a state-of-the-art specialist on a specific task [259]. In medicine, this principle suggests that generalist models should be compared with strong specialist or standard-of-care baselines for each clinically relevant task, using prespecified superiority or non-inferiority criteria appropriate to the intended use and clinical risk [42]. Cross-task validation additionally requires case-level coherence across outputs: a report claim should be compatible with localization, a risk estimate should not contradict the described disease burden, and a tool-using workflow should expose rather than conceal upstream processing failures.

**Representative methods and validation demands.** Evaluation in this family is often organized around benchmarks rather than individual models. M3D [117], MedVL-CT69K [81], Triad [199], MedVista3D [89], Merlin [3], and CuriaBench [200] probe broad volumetric task performance across

retrieval, reporting, segmentation, registration, downstream prediction, and foundation-model transfer. Grounding- and reasoning-focused benchmarks then examine whether model outputs remain consistent with the underlying volumetric evidence. CT-SpatialVQA [190] targets semantic-spatial CT reasoning; Med-StepBench [191] evaluates step-wise hallucination in oncological PET/CT; CORTEX [192] adds structured four-stage reasoning traces for chest CT; ReportQA [193] evaluates report utility through QA; and Oncology VQA [194] tests private 3D oncology cohorts with contamination-aware blind ablations.

Agent-era evaluation adds a process dimension. MedOpenClaw [134], MedVistaGym [135], ABRA [195], MedCTA [196], RadA-BenchPlat [197], and RadSaFE-200 [198] evaluate full-study interaction, viewer or tool use, agent-core behavior, trajectory fidelity, and unsafe-answer behavior.

These resources broaden evaluation beyond isolated tasks, but aggregate performance remains insufficient when outputs are assessed independently. Cross-task validation should examine case-level consistency and, for agentic systems, the validity of intermediate actions and acquired evidence. Final-answer accuracy alone cannot validate a workflow built on incorrect or unsupported intermediate steps.

**Current resources and remaining gaps.** Current benchmarks combine task-specific scores with measures of grounding, trace fidelity, workflow completion, and safety. These signals are usually reported separately, and few protocols determine whether an upstream error is propagated, detected, or corrected before it affects a downstream claim. The principal gap is therefore integrated validation across outputs and workflow stages. Contamination control, subgroup analysis, and external evaluation also remain limited, constraining claims of generality and clinical reliability. These limitations motivate the broader research agenda in Section 6.

The four application families operationalize the validation component of the Claim–Design–Validation alignment framework. Diagnostic claims require joint validation of spatial and language outputs (5.1). Prognostic claims require temporally defined, cohort-level evidence (5.2). Planning claims require end-to-end technical and workflow validation (5.3). Cross-task claims add the requirement that outputs and intermediate actions remain coherent (5.4). Across all four families, validation evidence often lags behind claim strength: many studies establish technical feasibility, fewer demonstrate external robustness, and only a limited number of studies evaluate workflow benefit or clinical utility.

# 6 Discussion and Future Directions

The preceding sections demonstrate that progress in this field has been driven not merely by model scaling, but by a fundamental sequence of representational and system-level shifts. We summarize this trajectory as a continuous progression: from 2D adaptation to native 3D foundation models, and ultimately toward agentic systems. This pathway aligns with the broader evolution of multimodal AI from lower-dimensional perception toward higher-dimensional, action-oriented capabilities [43, 46, 260].

This evolution began with a fundamental shift in how volumetric features are extracted. Early frameworks (e.g., LLaVA-Med [23], Med-Flamingo [27]) often adapted mature 2D priors, processing volumetric data as independent slices or reformatted views. While practically accessible, this slice-level adaptation risks losing through-plane continuity and long-range anatomical context. To bridge this gap, intermediate hybrid architectures (e.g., RadFM [106], Photon [108], and OmniCT [109]) have explored how far mature 2D priors can be reused while maintaining a pathway for volumetric context. Ultimately, however, the field is transitioning toward native 3D foundational representations (e.g., CT2Rep [115], Merlin [3], and Med3D-R1 [29]) to fully preserve spatial integrity. Yet, this full-scan extraction introduces a new structural bottleneck: the sheer volume of 3D data vastly exceeds the context limits of current LLMs, necessitating advanced compression and token-routing strategies.

To manage these context constraints and support the procedural complexity of volumetric interpretation, the field is increasingly extending toward agentic workflows. Rather than relying on a single forward pass, recent architectures (e.g., CT-Agent [35], RadAgent [37], MedScribe [133]) have decomposed complex radiology tasks into iterative, multi-step processes involving targeted volumetric inspection, dynamic tool invocation, and long-term memory updates. The core-corpus landscape in Figure 8 maps this transition along two complementary axes: temporal evolution and modality coverage. The temporal panel shows that early work established volumetric representations and evaluation resources, 3D and hybrid MLLMs

expanded rapidly after 2024, and advanced agentic workflow systems became concentrated in 2025–2026. The modality panel shows that the literature remains heavily CT-centric, while sequence-aware MRI models, PET/nuclear-medicine systems, and multi-modality or general 3D resources increasingly diversify the field. Read together, these patterns indicate that agentic architectures are a recent workflow-level extension of the 3D foundation-model literature rather than an independent replacement for it. They also show that 3D foundation models are transitioning from standalone solvers into modular components of traceable clinical pipelines. Thus, future validation frameworks should shift toward effective workflow integration, verifiable tool execution bounds, and collaborative human oversight.

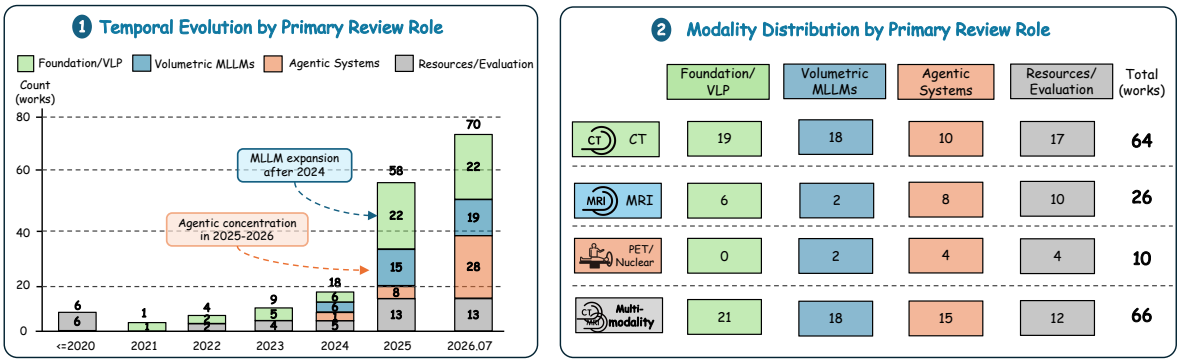


Figure 8: **Core-corpus landscape of volumetric radiology AI.** The figure summarizes the deduplicated core corpus of volumetric-radiology publications and resources included in the quantitative landscape analysis. Each work is assigned once to one of four primary review roles: foundation models/vision–language pre-training, volumetric radiology MLLMs, agentic systems, or clinical resources/evaluation. The right panel cross-tabulates the same assignments by dominant modality. General AI, review articles, and 2D-only background references are excluded.

Building upon these observations, Figure 9 summarizes our proposed research agenda. Rather than merely listing distinct challenges, this roadmap reflects a unified clinical trajectory: to systematically advance volumetric radiology, fundamental volumetric representations should be deeply coupled with scanning physics, while static multimodal interpretation should evolve into collaborative human-AI workflows. Ultimately, bridging the gap between digital diagnostic outputs and physical clinical reality will require agents to be grounded in world models, self-evolving technologies, and embodied physical infrastructure, all supported by rigorous, claim-specific data validation standards.

## 6.1 From 2D Surrogates to Volumetric and Acquisition-Aware Intelligence

Progress in 3D volumetric models depends on treating volumetric scans as native, physics-grounded structures rather than proxy modalities. The core modeling challenges involve managing informational density through clinically salient compression and unifying visual data with scanning physics metadata. This transition reflects the shifting trade-offs between computational tractability and clinical fidelity.

**Volumetric modeling should preserve native spatial continuity.** Volumetric radiology contains a combination of local details, global anatomy, and sparse, clinically decisive abnormalities. Therefore, native 3D architectures should account for this complexity rather than just scaling up 2D networks. Self-supervised methods already exploit this structure through volumetric restoration, masked modeling, spatial deformation, and geometry-aware contrastive objectives. Models such as Genesis [14], Swin UNETR [15], VoCo [16], and SPECTRE [63] show that volumetric continuity should be preserved as a fundamental representation principle, rather than being treated as a limited set of 2D surrogates.

**Compression policies should optimize token efficiency and be guided by clinical salience.** A major bottleneck in scaling medical VLMs from 2D to 3D is efficiently encoding massive, highly redundant volumetric data into a compact visual space. Because full-resolution 3D information is too large for current MLLM interfaces to process, achieving low-budget tokenization for language alignment is critical. Consequently, current systems should adopt specific compression policies for inputs or extracted visual tokens to balance performance and computational cost. Approaches like M$^3$D-LaMed [117] use perceiver-style pooling, while Med3DVLM [30], MedPruner [228], and Photon [108] make token efficiency an

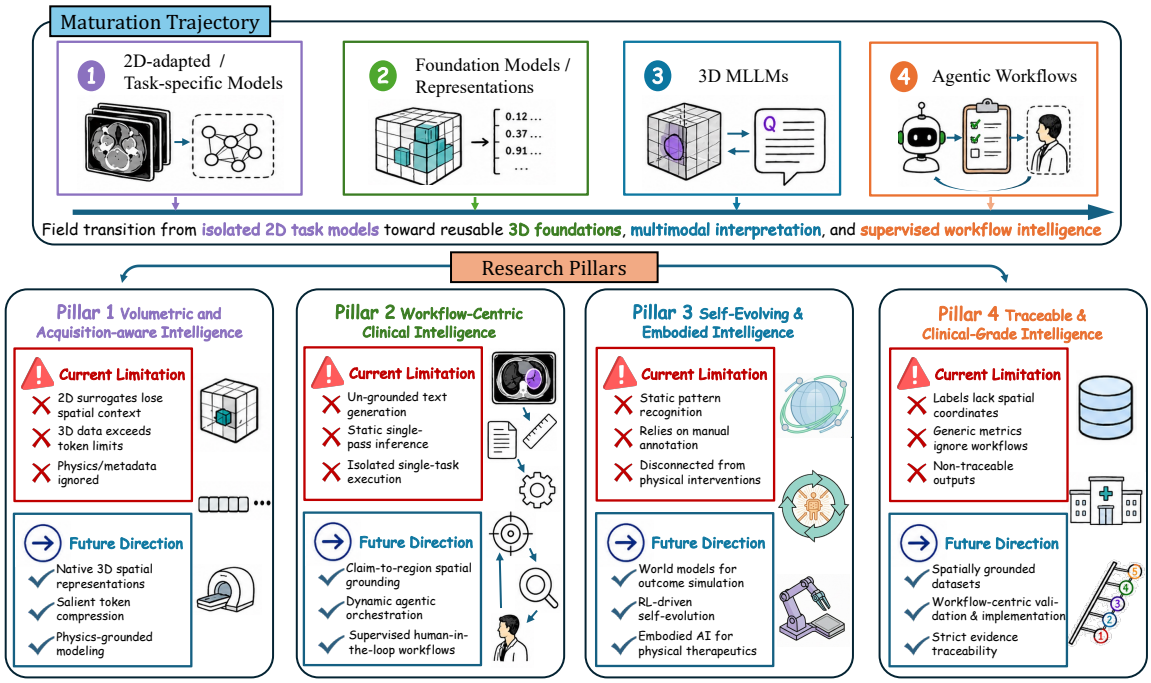


Figure 9: **A research agenda for volumetric radiology intelligence.** The top panel illustrates the field's transition from isolated, 2D-adapted task models toward reusable 3D foundations, multimodal interpretation, and supervised agentic workflows. To drive this evolution, the bottom panel outlines a research agenda structured across four core pillars of intelligence: (1) volumetric and acquisition-aware, (2) workflow-centric clinical, (3) self-evolving and embodied, and (4) traceable and clinical-grade, contrasting current limitations with future directions.

explicit design target. Future compression policies should be evaluated not only by aggregate task scores but also by their ability to selectively preserve small findings, spatial relations, and task-relevant regions, even when much of the background volume is discarded.

**Acquisition context should be explicitly represented alongside spatial anatomy.** CT phase, reconstruction kernel, MRI sequence, slice thickness, and scanner protocols are not incidental metadata; they are major sources of domain shift that alter the diagnostic meaning of visual patterns. Early methods point in this direction: MR-CLIP [76] uses DICOM metadata for semantic alignment, while $M^3$AE [59] learns from multi-sequence MRI inputs. Without integrating physics and protocol constraints into the representation space, models may produce plausible language while ignoring the underlying conditions that make a radiological finding valid and safe for clinical use.

## 6.2 From Static Perception to Workflow-centric Clinical Intelligence

Progress in volumetric radiology models will inevitably shift from static, perception-based interpretation toward operational, workflow-centric clinical intelligence. Looking forward, we anticipate a transition organized along three main axes: establishing token-level spatial grounding, expanding into multi-step agentic orchestration, and ultimately designing for bounded, human-AI symbiotic workflows.

**Interpretability should be grounded in image regions and clinical claims.** To advance beyond current limitations, future interpretability in volumetric radiology should transition from surface-level text generation to rigorous spatial grounding. Currently, a fluent report might lack volumetric support, and correct VQA answers often lack reliable localization. While fine-grained alignment methods, such as CT-GLIP [79], MG-3D [80], RadFinder [82], SCALE-VLP [96], MedRegion-CT [125], and discriminative-guided 3D CT grounding [223], have begun to decompose alignment across anatomy and report sentences, the critical next step is to enforce strict traceability. Future models should guarantee that every generated clinical claim can be systematically traced back to exact anatomical regions, precise measurements, or explicitly grounded visual evidence.

**Agents should evolve from passive inference to dynamic workflow orchestration.** Realizing true

clinical intelligence requires transforming models into operational agents rather than relying on static perception. Current MLLM inference remains a passive exercise, constrained by the assumption that all volumetric context is captured in a single forward pass. To overcome this, architectures should grant models the autonomy to navigate 3D scans, invoke clinical tools, and iteratively refine findings. Recent developments have begun introducing these multi-step capabilities, including CT-Agent [35] and 3DMedAgent [36] for volumetric QA; PD-CTAgent [136] for policy-constrained phase-sufficiency decisions; RadAgent [37], MedScribe [133], and Radiologist Copilot [39] for report generation; and CT-Flow [38] and MedOpenClaw [134] for interactive navigation. However, these methods currently focus on isolated tasks or retrospective evidence-acquisition settings. Looking ahead, research should expand these isolated, single-task systems to span the entire clinical continuum, from initial diagnosis to treatment planning and longitudinal patient management, ultimately delivering comprehensive care through structured, actionable operations [261] and interactive conversational support [262, 263]. As evidenced by the LungIMPACT trial [264], optimizing diagnostic steps in silos fails to resolve downstream bottlenecks without parallel expansions in subsequent pathway capacities. Therefore, future agents should be designed to orchestrate macro-workflows, dynamically harmonizing operations to optimize multi-disciplinary care delivery.

**Workflow intelligence should be anchored in supervised human-AI collaboration.** As medical agents evolve from static question-answering to dynamic, environment-aware conversational systems [262, 263, 265], their integration into volumetric radiology should reject unsupervised autonomy in favor of role-bounded collaboration. True intelligence requires systems that process multimodal inputs and yield inspectable results strictly under human jurisdiction. Methods like DosimeTron [155], MARTP [158], and SAGE [159] exemplify this synergy by navigating complex pipelines while deferring to clinicians during high uncertainty. Ultimately, the true measure of a clinical agent is not its capacity for independent execution, but its ability to safely amplify human expertise. By ensuring transparent reasoning and seamless handoffs, future agents will transcend mere automation to become trusted clinical copilots, bridging the gap between computational power and ultimate medical accountability.

### 6.3 From Digital Boundaries to Self-Evolving and Medical Embodied Intelligence

Progress in 3D medical AI should transcend static processing to embrace real-world interactions. To bridge digital diagnostics and physical interventions, research should augment clinical agents across three trajectories: integrating world models for dynamic environmental understanding, enabling continuous adaptation via self-evolving reinforcement, and grounding decisions through physical embodiment. Ultimately, these trajectories will transform agents from passive tools into active clinical participants.

**Environmental understanding should be expanded via clinical world models.** While current 3D MLLMs typically operate as static pattern discriminators, recent breakthroughs such as Brain-WM [266] and CLARITY [267] challenge this bottleneck by predicting sequential MRI trajectories and simulating patient outcomes. Synthesizing these advancements reveals a paradigm shift: environmental understanding requires expansion via clinical world models. Crucially, this modeling can extend beyond the visual domain and be fundamentally anchored in language. Inspired by general-purpose LLMs like Qwen-AgentWorld [268], which simulate complex interactive environments via next-state text prediction, clinical world models can leverage agentic LLMs to project patient trajectories. Under this setting, projecting these patient trajectories essentially constructs a simulated clinical world. Rather than acting as static predictive generators, these models serve as multi-modal sandboxes that execute a continuous "*perception-dynamics-planning*" loop [269, 270]. By internalizing this anticipatory meta-reasoning, agents transition from passive observers to active planners, mentally simulating post-interventional clinical notes and evaluating treatments prior to real-world execution.

**Sequential planning should progress from structured loop engineering to self-evolution.** Complex volumetric workflows, such as longitudinal tumor tracking and multimodal image analysis, require robust architectural foundations before cognitive scaling. System development in future radiology intelligence will naturally begin with loop engineering [209, 271, 272]. Such control loops decompose complex imaging tasks into executable and verifiable steps, such as anatomical localization, cross-modal registra-

tion, and diagnostic validation, while enabling agents to iteratively correct intermediate errors through feedback. Once this structured pipeline reliably generates high-quality diagnostic trajectories, agentic RL can be integrated as the cognitive layer to optimize these volumetric decision paths [273]. Beyond individual trajectory optimization, autoresearch enables agents to continuously propose, evaluate, and retain improvements to their planning strategies, tools, and workflow configurations [249, 274]. Together with agentic RL, this capability supports macro-level self-evolution. Successful clinical workflows can, for example, be distilled into reusable skill documents [275–277]. This structured developmental paradigm offers a highly scalable pathway to mitigate scanner-induced domain drift and expand clinical capabilities without continuous human annotation.

**Therapeutic execution should be grounded in embodied AI.** The ultimate capability scaling of clinical agents involves translating high-dimensional computational reasoning into physical action domains, effectively eliminating the boundary between passive diagnostic radiology and direct physical intervention [278]. In an embodied paradigm, agents will seamlessly link real-time imaging modalities, such as fluoroscopy and cone-beam CT, to autonomous navigation and robotic manipulation. Specific operational scenarios include utilizing multi-agent inference pipelines to analyze real-time angiographic sequences and autonomously guide endovascular catheters through complex vascular topographies [279]. Additionally, integrating intraoperative MRI and interventional ultrasound with robotic manipulators facilitates precise needle trajectory tracking for minimally invasive percutaneous procedures [280]. By executing closed-loop physical interventions [281], including image-guided targeted tissue biopsy, radiofrequency ablation, and adaptive radiotherapy beam alignment, embodied agents will anchor digital predictions into real-world physical therapeutics, realizing true end-to-end clinical intelligence.

## 6.4 From Benchmarks to Traceable and Clinical-Grade Intelligence

The transition of volumetric radiology AI into real-world workflows requires moving beyond technical feasibility. Achieving true clinical-grade intelligence relies on a holistic ecosystem comprising spatially grounded data, workflow-centric evaluation, and traceable clinical translation.

**Data infrastructure should co-evolve to support precise spatial grounding.** The shift from adapted 2D models to native 3D systems is bottlenecked by annotation granularity. Volumetric datasets like MedMNIST v2 [162] and AbdomenAtlas [166], alongside multimodal resources mapping scans to reports (e.g., CT-RATE [19] and $M^3D$ [117]), are foundational but capture only a fraction of clinical diversity. Crucially, weakly supervised image-report pairs risk introducing label noise if they lack exact anatomical coordinates. Future curation should prioritize spatially grounded annotations, exemplified by RadGenome-Chest CT [169], AutoRG-Brain [170], and SpatialMed [176]. These datasets explicitly tether textual outputs to volumetric evidence, forming the bedrock for reliable AI.

**Evaluation paradigms should shift from generic metrics to workflow-centric clinical validation.** Traditional isolated metrics (e.g., Dice and AUROC) solely focus on geometric or discriminative performance but fail to guarantee workflow safety or diagnostic triage readiness. Platforms like Merlin [3], MedOpenClaw [134], and RadSaFE-200 [198] are broadening capability assessments. Furthermore, as recent guidelines emphasize [282], evaluations should strictly align with the intended clinical function. A diagnostic tool requires rigorous calibration; a report generator demands factual consistency; and an agentic orchestrator should be judged on tool validity and human handoff efficacy [46]. Performance should be measured by fitness for the intended clinical workflow rather than benchmark dominance.

**Clinical translation should employ a staged validation ladder and strict evidence traceability.** Bridging the gap from targeted benchmarks to actual deployment requires a phased validation approach. This progression should begin with retrospective testing, advance through reader studies and workflow simulations, and culminate in prospective clinical trials and continuous post-deployment monitoring. While existing frameworks like MI-CLAIM [283], CONSORT-AI [284], DECIDE-AI [285], CLAIM 2024 [286], STARD-AI [287], and FUTURE-AI [41] formalize this progression, they should be specifically adapted for volumetric radiology by incorporating image-grounded spatial traces. Furthermore, executing this pipeline safely demands granular evidence traceability. General risks such as dataset shift [288], algorithmic bias [289], data poisoning [290], and cognitive bias [291] are heavily compounded

by imaging-specific vulnerabilities like scanner protocol variability and reconstruction artifacts. To mitigate these multifaceted threats, models should explicitly link every clinical output to the exact study, series, slice range, or prior examination that supported it. This transparent traceability forms the indispensable basis for debugging, safe human review, and institutional governance.

**Implementation science should be integrated across the clinical translation pathway.** Implementation science studies the methods and contextual determinants that shape how evidence-based interventions are adopted, integrated, scaled, and sustained in routine practice. Within the AI translation pathway, technical and clinical validation assess whether a system performs as intended under defined conditions, whereas implementation science examines how that system is incorporated into a specific workflow, organization, and user environment. It therefore complements validation by addressing infrastructure, usability, human–AI interaction, organizational readiness, trust, and other determinants of routine use [292]. For volumetric radiology, hybrid effectiveness–implementation studies, which evaluate clinical effectiveness and implementation outcomes within the same study, should assess adoption, fidelity, scalability, and sustainability alongside task-specific performance and clinical outcomes, with stakeholders engaged from early development through post-deployment monitoring. In precision oncology and theranostics, the augmented-oncologist concept illustrates this perspective: implementation-ready AI and patient-specific digital twins may support clinician-led diagnosis, treatment planning, education, and longitudinal decision-making [293]. Together, these considerations extend the bounded human–AI collaboration described in Section 6.2 and frame implementation as an iterative process across the clinical translation pathway.

### 6.5 Scope and Limitations

This review focuses specifically on 3D volumetric radiology, including CT, MRI, PET, single-photon emission computed tomography (SPECT), and hybrid acquisitions, and on the technical progression from representation learning to foundation models, MLLMs, agents, evaluation, and applications. It does not attempt to provide a complete survey of all 2D medical imaging models, all non-imaging medical LLMs, or all legal, economic, and reimbursement issues related to clinical deployment. Furthermore, the literature window for this review closes in July 2026. Because the agentic literature is developing rapidly, several systems discussed are available as preprints or early reports. We therefore interpret them as indicators of emerging design patterns rather than as proof of deployment readiness.

## 7 Conclusion

Volumetric radiology AI is progressing from 2D-adapted task models toward native volumetric representations and agentic workflows. This progression comprises a representational advance that preserves clinically decisive volumetric information and aligns it with clinical language, and an operational advance that enables this evidence to be acquired, verified, and coordinated beyond a single forward pass. Selected 2D views remain effective when decisive evidence is confined to a limited set of images, whereas native volumetric modeling becomes essential when interpretation depends on full-volume spatial relationships. Agentic systems extend this foundation by coupling volumetric perception with tools, memory, and iterative workflow orchestration.

Clinical credibility depends on alignment among the intended claim, the evidence preserved by the model, the actions performed by the system, and the validation supporting that claim. The Claim–Design–Validation framework formalizes this alignment and identifies traceability and clearly defined human oversight as cross-cutting requirements. Future progress will depend on stronger native volumetric data resources, more reliable grounding and validation, and tighter integration with clinical tools and longitudinal context. As these components mature, volumetric radiology systems may evolve into traceable, human-supervised systems that support radiologists across diagnostic, planning, and longitudinal workflows.

# References


[1] Jung-Oh Lee, Hong-Yu Zhou, Tyler M Berzin, Daniel K Sodickson, and Pranav Rajpurkar. Multimodal generative AI for interpreting 3D medical images and videos. *npj Digital Medicine*, 8(1):273, 2025.

[2] Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical AI. *Nature Medicine*, 28(9):1773–1784, 2022.

[3] Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, et al. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, 652(8112):1–11, 2026.

[4] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[5] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, 2024.

[6] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer learning for 3D medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

[7] Brendan S Kelly, Conor Judge, Stephanie M Bollard, Simon M Clifford, Gerard M Healy, Awsam Aziz, Prateek Mathur, Shah Islam, Kristen W Yeom, Aonghus Lawlor, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). *European Radiology*, 32(11):7998–8007, 2022.

[8] Rushabh Doshi, Kanhai S Amin, Pavan Khosla, Simar S Bajaj, Sophie Chheang, and Howard P Forman. Quantitative evaluation of large language models to streamline radiology report impressions: A multimodal retrospective analysis. *Radiology*, 310(3):e231593, 2024.

[9] Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3D-RAD: A comprehensive 3D radiology med-VQA dataset with multi-temporal analysis and diverse diagnostic tasks. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[10] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM Van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, 48(4):441–446, 2012.

[11] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016.

[12] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017.

[13] Ozgun Cicek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016.

[14] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical Image Analysis*, 67:101840, 2021.

[15] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of Swin transformers for 3D medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022.

[16] Linshan Wu, Jiaxin Zhuang, and Hao Chen. Large-scale 3D medical image pre-training with geometric context priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3801–3818, 2025.

[17] Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng. Self-supervised feature learning for 3D medical images by playing a rubik's cube. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–428. Springer, 2019.

[18] Jiuwen Zhu, Yuexiang Li, Yifan Hu, Kai Ma, S Kevin Zhou, and Yefeng Zheng. Rubik's Cube+: A self-supervised feature learning framework for 3d medical image analysis. *Medical image analysis*, 64:101746, 2020.

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

[20] Cameron Beeche, Joonghyun Kim, Hamed Tavolinejad, Bingxin Zhao, Rakesh Sharma, Jeffrey Duda, James Gee, Farouk Dako, Anurag Verma, Colleen Morse, et al. A pan-organ vision-language model for generalizable 3D CT representations. *medRxiv*, 2025.

[21] Vishwanatha M Rao, Michael Hla, Michael Moor, Subathra Adithan, Stephen Kwak, Eric J Topol, and Pranav Rajpurkar. Multimodal generative AI for medical image interpretation. *Nature*, 639(8056):888–896, 2025.

[22] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616 (7956):259–265, 2023.

[23] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pages 28541–28564, 2023.

[24] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

[25] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal LLMs at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370, 2024.

[26] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-R1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE Transactions on Medical Imaging*, 45(6):2727–2737, 2026.

[27] 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 *Machine Learning for Health (ML4H)*, volume 225, pages 353–367. PMLR, 2023.

[28] Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-Med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

[29] Haoran Lai, Zihang Jiang, Kun Zhang, Qingsong Yao, Rongsheng Wang, Zhiyang He, Xiaodong Tao, Wei Wei, and Shaohua Kevin Zhou. Med3D-R1: Incentivizing clinical reasoning in 3D medical vision-language models for abnormality diagnosis. *arXiv preprint arXiv:2602.01200*, 2026.

[30] Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3DVLM: An efficient vision-language model for 3D medical image analysis. *IEEE Journal of Biomedical and Health Informatics*, 30(3):2524–2536, 2025.

[31] Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. MedGemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

[32] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, pages 79410–79452, 2024.

[33] Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. MMedAgent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8745–8760, 2024.

[34] Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. MedAgent-Pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*, 2025.

[35] Yuren Mao, Wenyi Xu, Yuyang Qin, and Yunjun Gao. CT-Agent: a multimodal-LLM agent for 3D CT radiology question answering. *arXiv preprint arXiv:2505.16229*, 2025.

[36] Ziyue Wang, Linghan Cai, Chang Han Low, Haofeng Liu, Junde Wu, Jingyu Wang, Rui Wang, Lei Song, Jiang Bian, Jingjing Fu, et al. 3DMedAgent: Unified perception-to-understanding for 3D medical analysis. *arXiv preprint arXiv:2602.18064*, 2026.

[37] Mélanie Roschewitz, Kenneth Styppa, Yitian Tao, Jiwoong Sohn, Jean-Benoit Delbrouck, Benjamin Gundersen, Nicolas Deperrois, Christian Bluethgen, Julia Vogt, Bjoern Menze, et al. RadAgent: A tool-using AI agent for stepwise interpretation of chest computed tomography. *arXiv preprint arXiv:2604.15231*, 2026.

[38] Yannian Gu, Xizhuo Zhang, Linjie Mu, Yongrui Yu, Zhongzhen Huang, Shaoting Zhang, and Xiaofan Zhang. CT-Flow: Orchestrating CT interpretation workflow with model context protocol servers. *arXiv preprint arXiv:2603.00123*, 2026.

[39] Yongrui Yu, Zhongzhen Huang, Linjie Mu, Shaoting Zhang, and Xiaofan Zhang. Radiologist Copilot: An agentic assistant with orchestrated tools for radiology reporting with quality control. *arXiv preprint arXiv:2512.02814*, 2025.

[40] Sune Holm, Daria Ferrara, Miriam Pepponi, Elisabetta Abenavoli, Armin Frille, Shaul Duke, Stefan Grünert, Marcus Hacker, Bengt Hennig, Swen Hesse, et al. Explainable AI in nuclear medicine. *European Journal of Nuclear Medicine and Molecular Imaging*, 53(4):2648–2651, 2026.

[41] Karim Lekadir, Alejandro F Frangi, Antonio R Porras, Ben Glocker, Celia Cintas, Curtis P Langlotz, Eva Weicken, Folkert W Asselbergs, Fred Prior, Gary S Collins, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, 388:e081554, 2025.

[42] Babak Saboury, Tyler Bradshaw, Ronald Boellaard, Irène Buvat, Joyita Dutta, Mathieu Hatt, Abhinav K Jha, Quanzheng Li, Chi Liu, Helena McMeekin, et al. Artificial intelligence in nuclear medicine: Opportunities, challenges, and responsibilities toward a trustworthy ecosystem. *Journal of Nuclear Medicine*, 64(2):188–196, 2023.

[43] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion*, 117:102888, 2025.

[44] Hanguang Xiao, Ningzhi Hui, Yong Xu, Zhipeng Li, and Jincheng Peng. Medical multimodal large language models: A survey. *Information Fusion*, 134:104386, 2026.

[45] Jing Wu, Yuli Wang, Zhusi Zhong, Weihua Liao, Natalia Trayanova, Zhicheng Jiao, and Harrison X Bai. Vision-language foundation model for 3D medical imaging. *npj Artificial Intelligence*, 1(1):17, 2025.

[46] Christian Bluethgen, Dave Van Veen, Daniel Truhn, Jakob Nikolas Kather, Michael Moor, Malgorzata Polacin, Akshay Chaudhari, Thomas Frauenfelder, Curtis P. Langlotz, Michael Krauthammer, and Farhad Nooralahzadeh. Agentic systems in radiology: Design, applications, evaluation, and challenges. *arXiv preprint arXiv:2510.09404*, 2025.

[47] Jeya Maria Jose Valanarasu, Yucheng Tang, Dong Yang, Ziyue Xu, Can Zhao, Wenqi Li, Vishal M Patel, Bennett Allan Landman, Daguang Xu, Yufan He, et al. Disruptive autoencoders: Leveraging low-level features for 3D medical image pre-training. In *Medical Imaging with Deep Learning*, volume 250, pages 1553–1570. PMLR, 2024.

[48] Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Mariann Micsinai Balan, and Kevin Brown. Masked image modeling advances 3D medical image analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1970–1980, 2023.

[49] Jin Lee, Vu Dang, Gwang-Hyun Yu, Anh Le, Zahid Rahman, Jin-Ho Jang, Heonzoo Lee, Kun-Yung Kim, Jin-Sul Kim, and Jin-Young Kim. HU-based foreground masking for 3D medical masked image modeling. In *International Workshop on Applications of Medical AI*, pages 122–131. Springer, 2025.

[50] Zhaohu Xing, Lei Zhu, Lequan Yu, Zhiheng Xing, and Liang Wan. Hybrid masked image modeling for 3D medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 28(4):2115–2125, 2024.

[51] Jiaxin Zhuang, Linshan Wu, Qiong Wang, Peng Fei, Varut Vardhanabhuti, Lin Luo, and Hao Chen. MiM: Mask in mask self-supervised pre-training for 3D medical image analysis. *IEEE Transactions on Medical Imaging*, 44(9):3727–3740, 2025.

[52] Junyan Lyu, Perry F Bartlett, Fatima A Nasrallah, and Xiaoying Tang. Masked deformation modeling for volumetric brain MRI self-supervised pre-training. *IEEE Transactions on Medical Imaging*, 44(3): 1596–1607, 2024.

[53] Yuting He, Guanyu Yang, Rongjun Ge, Yang Chen, Jean-Louis Coatrieux, Boyu Wang, and Shuo Li. Geometric visual similarity learning in 3D medical image self-supervised pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9538–9547, 2023.

[54] Liam Chalcroft, Jenny Crinion, Cathy J. Price, and John Ashburner. Unified 3D MRI representations via sequence-invariant contrastive learning. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 63–74. Springer, 2025.

[55] Sheng Wang, Zihao Zhao, Zhenrong Shen, Bin Wang, Qian Wang, and Dinggang Shen. Improving self-supervised medical image pre-training by early alignment with human eye gaze information. *IEEE Transactions on Medical Imaging*, 44(10):4063–4072, 2025.

[56] Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. UniMiSS: Universal medical self-supervised learning via breaking dimensionality barrier. In *European Conference on Computer Vision*, pages 558–575. Springer, 2022.

[57] Fei Gao, Siwen Wang, Fandong Zhang, Hong-Yu Zhou, Yizhou Wang, Churan Wang, Gang Yu, and Yizhou Yu. Cross-dimensional medical self-supervised representation learning based on a pseudo-3D transformation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 178–188. Springer, 2024.

[58] Hong-Yu Zhou, Chixiang Lu, Chaoqi Chen, Sibei Yang, and Yizhou Yu. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8020–8035, 2023.

[59] Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. M3AE: Multimodal representation learning for brain tumor segmentation with missing modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1657–1665, 2023.

[60] Shaohao Rui, Lingzhi Chen, Zhenyu Tang, Lilong Wang, Mianxin Liu, Shaoting Zhang, and Xiaosong Wang. Multi-modal vision pre-training for medical image analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5164–5174, 2025.

[61] Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, Qi Wu, and Yong Xia. Continual self-supervised learning: Towards universal multi-modal medical data representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11114–11124, 2024.

[62] Weicheng Zhu, Haoxu Huang, Huanze Tang, Rushabh Musthyala, Boyang Yu, Long Chen, Emilio Vega, Thomas O'Donnell, Seena Dehkharghani, Jennifer A Frontera, et al. 3D foundation model for generalizable disease detection in head computed tomography. *arXiv preprint arXiv:2502.02779*, 2025.

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

[64] Tony Xu, Sepehr Hosseini, Chris Anderson, Anthony Rinaldi, Rahul G Krishnan, Anne L Martel, and Maged Goubran. A generalizable 3D framework and model for self-supervised learning in medical imaging. *npj Digital Medicine*, 8(1):639, 2025.

[65] Ioannis Gatopoulos, Nicolas Känzig, Sebastian Otálora, and Fei Tang. CoralBay: A self-supervised CT foundation model. *arXiv preprint arXiv:2606.03888*, 2026.

[66] Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Chang-bae Bang, Lin Zhao, Wanyi Fu, Jinglei Lv, Jorge Sepulcre, Carl Yang, et al. Towards a general-purpose foundation model for functional mri analysis. *Nature Biomedical Engineering*, pages 1–12, 2026.

[67] Hui Zhao, Ruipeng Zhang, Zhiyu Wang, Yifeng Gu, Shengyuan Xu, Sheng Wang, and Yuehua Li. Bonecot: multicentre validation of a whole-body skeleton foundation model for bone metastases guided by clinician-derived chain of thought. *Nature Biomedical Engineering*, pages 1–13, 2026.

[68] Zijian Dong, Yi Lin, Ji Fang, Jianxiong Zhou, Kwun Kei Ng, and Juan Helen Zhou. Brainfibre: A foundation model via information decomposition for brain microstructure. *arXiv preprint arXiv:2607.00573*, 2026.

[69] Qi Chen, Shuhan Ding, Yu Gu, Nan Liu, Jiang Bian, Alan Yuille, Zongwei Zhou, and Jingjing Fu. Foundation vae for ct reconstruction, augmentation, and generation. In *Forty-third International Conference on Machine Learning*, 2026.

[70] 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*, pages 525–536. Springer, 2023.

[71] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, 2023.

[72] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, 2022.

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

[74] Che Liu, Cheng Ouyang, Yinda Chen, César Quilodrán-Casas, Lei Ma, Jie Fu, Yike Guo, Anand Shah, Wenjia Bai, and Rossella Arcucci. T3D: Advancing 3D medical vision-language pre-training by learning multi-view visual consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6704–6714, 2025.

[75] Yingtai Li, Haoran Lai, Xiaoqian Zhou, Shuai Ming, Wenxin Ma, Wei Wei, and Shaohua Kevin Zhou. More performant and scalable: Rethinking contrastive vision-language pre-training of radiology in the LLM era. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 348–357. Springer, 2025.

[76] Mehmet Yigit Avci, Pedro Borges, Virginia Fernandez, Paul Wright, Mehmet Yigitsoy, Sebastien Ourselin, and Jorge Cardoso. Metadata-aligned 3D MRI representations for contrast understanding and quality control. *arXiv preprint arXiv:2511.00681*, 2025.

[77] Yuhui Tao, Zhongwei Zhao, Zilong Wang, Xufang Luo, Feng Chen, Kang Wang, Chuanfu Wu, Xue Zhang, Shaoting Zhang, Jiaxi Yao, et al. A disease-centric vision-language foundation model for precision oncology in kidney cancer. *Nature Communications*, 2026.

[78] Bowen Shi, Weiwei Cao, Ruifeng Yuan, Wanxing Chang, Wenrui Dai, Hongkai Xiong, Ling Zhang, and Jianpeng Zhang. Disease-centric vision-language pretraining with hybrid visual encoding for 3D computed tomography. *arXiv preprint arXiv:2606.25546*, 2026.

[79] Jingyang Lin, Yingda Xia, Jianpeng Zhang, Ke Yan, Kai Cao, Le Lu, Jiebo Luo, and Ling Zhang. CT-GLIP: 3D grounded language-image pretraining with CT scans and radiology reports for full-body scenarios. *arXiv preprint arXiv:2404.15272*, 2024.

[80] Xuefeng Ni, Linshan Wu, Jiaxin Zhuang, Qiong Wang, Mingxiang Wu, Varut Vardhanabhuti, Lihai Zhang, Hanyu Gao, and Hao Chen. MG-3D: Multi-grained knowledge-enhanced 3D medical vision-language pre-training. *arXiv preprint arXiv:2412.05876*, 2024.

[81] Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, et al. Large-scale and fine-grained vision-language pre-training for enhanced CT image understanding. *arXiv preprint arXiv:2501.14548*, 2025.

[82] Simon Ging, Philipp Arnold, Sebastian Walter, Hani Alnahas, Hannah Bast, Elmar Kotter, Jiancheng Yang, Behzad Bozorgtabar, and Thomas Brox. Learning to read where to look: Disease-aware vision-language pretraining for 3D CT. *arXiv preprint arXiv:2603.02026*, 2026.

[83] Jonggwon Park, Kyoyun Choi, Byungmu Yoon, Hong Geun Cho, and Bumcheol Hwang. RadZero3D: Bridging self-supervised video models and medical vision-language alignment for zero-shot chest CT interpretation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6742–6749, 2025.

[84] Jiayi Wang, Hadrien Reynaud, Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Bjoern Menze, and Bernhard Kainz. SigVLP: Sigmoid volume-language pre-training for self-supervised CT-volume adaptive representation learning. *arXiv preprint arXiv:2602.21735*, 2026.

[85] Rongsheng Wang, Fenghe Tang, Zihang Jiang, Yingtai Li, Xu Zhang, Haoran Lai, Wenxin Ma, Wei Wei, Zhiyang He, Xiaodong Tao, et al. ASAP: Advancing medical volumetric representation learning with anatomy-aware semantically-adaptive pre-training. *arXiv preprint arXiv:2606.00602*, 2026.

[86] Jonggwon Park, Seongeun Lee, Junhyun Park, Hannah Yun, Hyunwoong Kim, Sohyun Jeong, Hyewon Kang, Byungmu Yoon, and Kyoyun Choi. GLINT: Sparsely gated vision-language alignment for fine-grained radiology representations. *arXiv preprint arXiv:2606.03180*, 2026.

[87] Julien Khlaut, Charles Corbière, Baptiste Callard, Amaury Prat, Leo Butsanets, Antoine Saporta, Théo Danielou, Leo Machado, Korentin Le Floch, Tom Boeken, et al. Jolia: Concept-level vision-language alignment for 3D CT contrastive learning. *arXiv preprint arXiv:2606.24570*, 2026.

[88] Shuo Jiang, Yuhao Hong, Chunbo Jiang, Weihong Chen, Huangwei Chen, Shenghao Zhu, Beining Wu, Mingxuan Liu, Zhu Zhu, Feiwei Qin, et al. GLeVE: Graph-guided lesion grounding with proposal verification in 3D CT. *arXiv preprint arXiv:2605.22619*, 2026.

[89] Yuheng Li, Yenho Chen, Yuxiang Lai, Jike Zhong, Vanessa Wildman, and Xiaofeng Yang. MedVista3D: Vision-language modeling for reducing diagnostic errors in 3D CT disease detection, understanding and reporting. *arXiv preprint arXiv:2509.03800*, 2025.

[90] Yuheng Li, Yuxiang Lai, Maria Thor, Deborah Marshall, Zachary Buchwald, David S Yu, and Xiaofeng Yang. Towards universal text-driven CT image segmentation. *arXiv preprint arXiv:2503.06030*, 2025.

[91] Jiayu Lei, Lisong Dai, Haoyun Jiang, Chaoyi Wu, Xiaoman Zhang, Yao Zhang, Jiangchao Yao, Weidi Xie, Yanyong Zhang, Yuehua Li, et al. UniBrain: Universal brain MRI diagnosis with hierarchical knowledge-enhanced pre-training. *Computerized Medical Imaging and Graphics*, 122:102516, 2025.

[92] Ziyang Zhang, Yang Yu, Xulei Yang, and Si Yong Yeo. VELVET-Med: Vision and efficient language pre-training for volumetric imaging tasks in medicine. *arXiv preprint arXiv:2508.12108*, 2025.

[93] Chenhui Zhao, Yiwei Lyu, Asadur Chowdury, Edward Harake, Akhil Kondepudi, Akshay Rao, Xinhai Hou, Honglak Lee, and Todd Hollon. Towards scalable language-image pre-training for 3D medical imaging. *arXiv preprint arXiv:2505.21862*, 2025.

[94] Yuxiang Nie, Sunan He, Yequan Bie, Yihui Wang, Zhixuan Chen, Shu Yang, and Hao Chen. ConceptCLIP: Towards trustworthy medical AI via concept-enhanced contrastive langauge-image pre-training. *arXiv e-prints*, pages arXiv–2501, 2025.

[95] Yiming Lei, Zilong Li, Yan Shen, Junping Zhang, and Hongming Shan. CLIP-Lung: Textual knowledge-guided lung nodule malignancy prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 403–412. Springer, 2023.

[96] Ailar Mahdizadeh, Puria Azadi Moghadam, Xiangteng He, Shahriar Mirabbasi, Panos Nasiopoulos, and Leonid Sigal. SCALE-VLP: Soft-weighted contrastive volumetric vision-language pre-training with spatial-knowledge semantics. *arXiv preprint arXiv:2511.02996*, 2025.

[97] Haoran Lai, Zihang Jiang, Qingsong Yao, Rongsheng Wang, Zhiyang He, Xiaodong Tao, Weifu Lv, Wei Wei, and Shaohua Kevin Zhou. Bridged semantic alignment for zero-shot 3D medical image diagnosis. *IEEE Journal of Biomedical and Health Informatics*, pages 1–14, 2025.

[98] Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Qi Zhang, Tingbo Liang, et al. Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23041–23050, 2025.

[99] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

[100] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical AI. *NEJM AI*, 1(3): AIoa2300138, 2024.

[101] Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. Qilin-Med-VL: Towards Chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*, 2023.

[102] Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Song Xiaohui, et al. HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *International Conference on Machine Learning*, volume 267, pages 37975–37995. PMLR, 2025.

[103] Wei Dai, Peilin Chen, Chanakya Ekbote, and Paul Pu Liang. QoQ-Med: Building multimodal clinical foundation models with domain-aware GRPO training. *arXiv preprint arXiv:2506.00711*, 2025.

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

[105] Timothy Ossowski, Sheng Zhang, Qianchu Liu, Guanghui Qin, Reuben Tan, Tristan Naumann, Junjie Hu, and Hoifung Poon. OctoMed: Data recipes for state-of-the-art multimodal medical reasoning. *arXiv preprint arXiv:2511.23269*, 2025.

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

[107] Yiming Shi, Xun Zhu, Kaiwen Wang, Ying Hu, Chenyi Guo, Miao Li, and Ji Wu. Med-2E3: A 2D-enhanced 3D medical multimodal large language model. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2754–2759. IEEE, 2025.

[108] Chengyu Fang, Heng Guo, Zheng Jiang, Chunming He, Xiu Li, and Minfeng Xu. Photon: Speedup volume understanding with efficient multimodal large language models. *arXiv preprint arXiv:2603.25155*, 2026.

[109] Tianwei Lin, Zhongwei Qiu, Wenqiao Zhang, Jiang Liu, Yihan Xie, Mingjian Gao, Zhenxuan Fan, Zhaocheng Li, Sijing Li, Zhongle Xie, et al. OmniCT: Towards a unified slice-volume LVLM for comprehensive CT analysis. *arXiv preprint arXiv:2602.16110*, 2026.

[110] Jiayu Lei, Ziqing Fan, Yanyong Zhang, Weidi Xie, Ya Zhang, and Yanfeng Wang. Versatile vision-language model for 3D computed tomography. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(8): 5945–5954, 2026.

[111] Yiming Shi, Shaoshuai Yang, Xun Zhu, Haoyu Wang, Xiangling Fu, Miao Li, and Ji Wu. MedM-VL: What makes a good medical LVLM? In *International Workshop on Agentic AI for Medicine*, pages 290–299. Springer, 2025.

[112] Jianqin Liu, Weiwei Cao, Wanxing Chang, Ruifeng Yuan, Bowen Shi, Zhilin Zheng, Xianjie Zhang, Ling Zhang, Peng Wang, and Jianpeng Zhang. Radsight: Towards perceptually reliable multimodal radiology image understanding. *arXiv preprint arXiv:2607.22293*, 2026.

[113] Hangjie Yuan, Yichen Qian, Zhiwei Tang, Xianzhe Xu, Lirong Wu, Sicheng Yang, Jinwang Wang, Pengju Wang, Zhitao Zeng, Yizeng Han, et al. Clinfusion: A vision-centric multimodal llm system for holistic medical understanding. *arXiv preprint arXiv:2607.24743*, 2026.

[114] Tengfei Zhang, Ziheng Zhao, Xiaoman Zhang, Lisong Dai, Pengcheng Qiu, Ya Zhang, Yanfeng Wang, and Weidi Xie. A vision-language framework for comparative reasoning in radiology. *arXiv preprint arXiv:2606.06407*, 2026.

[115] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2Rep: Automated radiology report generation for 3D medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 476–486. Springer, 2024.

[116] Zhixuan Chen, Luyang Luo, Yequan Bie, and Hao Chen. Dia-LLaMA: Towards large language model-driven CT report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 141–151. Springer, 2025.

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

[118] 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, et al. Towards a holistic framework for multimodal large language models in three-dimensional brain CT report generation. *arXiv preprint arXiv:2407.02235*, 2024.

[119] Hao Chen, Wei Zhao, Yingli Li, Tianyang Zhong, Yisong Wang, Youlan Shang, Lei Guo, Junwei Han, Tianming Liu, Jun Liu, et al. 3D-CT-GPT: Generating 3D radiology reports through integration of large vision-language models. *arXiv preprint arXiv:2409.19330*, 2024.

[120] Haoran Lai, Zihang Jiang, Qingsong Yao, Rongsheng Wang, Zhiyang He, Xiaodong Tao, Wei Wei, Weifu Lv, and S Kevin Zhou. E3D-GPT: enhanced 3D visual foundation for medical vision-language model. *arXiv preprint arXiv:2410.14200*, 2024.

[121] Zhixuan Chen, Yequan Bie, Haibo Jin, and Hao Chen. Large language model with region-guided referring and grounding for CT report generation. *IEEE Transactions on Medical Imaging*, 44(8):3139–3150, 2025.

[122] Changsun Lee, Sangjoon Park, Cheong-Il Shin, Woo Hee Choi, Hyun Jeong Park, Jeong Eun Lee, and Jong Chul Ye. Read like a radiologist: efficient vision-language model for 3D medical imaging interpretation. *arXiv preprint arXiv:2412.13558*, 2024.

[123] Xiaodan Zhang, Yanzhao Shi, Junzhong Ji, Chengxin Zheng, and Liangqiong Qu. MEPNet: Medical entity-balanced prompting network for brain CT report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25940–25948, 2025.

[124] Yanzhao Shi, Xiaodan Zhang, Junzhong Ji, Haoning Jiang, Chengxin Zheng, Yinong Wang, and Liangqiong Qu. HSENet: Hybrid spatial encoding network for 3D medical vision-language understanding. *arXiv preprint arXiv:2506.09634*, 2025.

[125] Sunggu Kyung, Jinyoung Seo, Hyunseok Lim, Dongyeong Kim, Hyungbin Park, Jimin Sung, Jihyun Kim, Wooyoung Jo, Yoojin Nam, and Namkug Kim. MedRegion-CT: region-focused multimodal LLM for comprehensive 3D CT report generation. *arXiv preprint arXiv:2506.23102*, 2025.

[126] Arvind Murari Vepa, Yannan Yu, Jingru Gan, Anthony Cuturrufo, Weikai Li, Wei Wang, Fabien Scalzo, and Yizhou Sun. A multimodal LLM approach for visual question answering on multiparametric 3D brain MRI. *arXiv preprint arXiv:2509.25889*, 2025.

[127] Danyal Maqbool, Changhee Lee, Zachary Huemann, Samuel D Church, Matthew E Larson, Scott B Perlman, Tomas A Romero, Joshua D Warner, Meghan Lubner, Xin Tie, et al. PETAR: Localized findings generation with mask-aware vision-language modeling for PET automated reporting. *arXiv preprint arXiv:2510.27680*, 2025.

[128] Wenpei Jiao, Kun Shang, Hui Li, Ke Yan, Jiajin Zhang, Guangjie Yang, Lijuan Guo, Yan Wan, Xing Yang, Dakai Jin, et al. Vision-language models for automated 3D PET/CT report generation. *arXiv preprint arXiv:2511.20145*, 2025.

[129] Mariano Barone, Francesco Di Serio, Giuseppe Riccio, Antonio Romano, Marco Postiglione, Antonino Ferraro, and Vincenzo Moscato. Brain3D: Brain report automation via inflated vision transformers in 3D. *arXiv preprint arXiv:2602.22098*, 2026.

[130] Vanshali Sharma, Andrea M Bejar, Halil Ertugrul Aktas, Quoc-Huy Trinh, Debesh Jha, Gorkem Durak, and Ulas Bagci. Revisiting LLM adaptation for 3D CT report generation: A study of scaling and diagnostic priors. *arXiv preprint arXiv:2606.17213*, 2026.

[131] Xinran Li, Junlin Yang, Annabella Shewarega, Zongwei Zhou, Julius Chapiro, James S Duncan, and Lawrence H Staib. MRI2Rep: Autoregressive structured report generation for 3D liver MRI. *arXiv preprint arXiv:2606.25279*, 2026.

[132] Sijing Li, Zhongwei Qiu, Zhuoya Wang, Boxiang Yun, Zhenyu Yi, Jianwei Xu, Wenqiao Zhang, Yingda Xia, and Ling Zhang. E-MRL: Cross-view aligned evidence-driven multimodal reinforcement learning for reliable 3D tumor analysis. *arXiv preprint arXiv:2606.23888*, 2026.

[133] Giuseppe A Orlando, Paolo Papotti, Maria A Zuluaga, Olivier Humbert, and Marco Lorenzi. MedScribe: Clinically grounded CT reporting through agentic workflows. *arXiv preprint arXiv:2605.01779*, 2026.

[134] Weixiang Shen, Yanzhu Hu, Che Liu, Junde Wu, Jiayuan Zhu, Chengzhi Shen, Min Xu, Yueming Jin, Benedikt Wiestler, Daniel Rueckert, et al. MedOpenClaw: Auditable medical imaging agents reasoning over uncurated full studies. *arXiv preprint arXiv:2603.24649*, 2026.

[135] Meng Lu, Yuxing Lu, Yuchen Zhuang, Megan Mullins, Yang Xie, Guanghua Xiao, Charles Fleming, Wenqi Shi, and Xuan Wang. MEDVISTAGYM: A scalable training environment for thinking with medical images via tool-integrated reinforcement learning. *arXiv preprint arXiv:2601.07107*, 2026.

[136] Yanmeng Dong, Han Li, Yujia Li, Jingsong Liu, Xun Ma, Yanzhu Hu, Zhengyang Xu, Zhicheng Li, Nassir Navab, and Shaohua Kevin Zhou. Policy-driven ct-agent: Modeling phase-aware diagnostic control for clinically consistent ct reasoning. *arXiv preprint arXiv:2607.10748*, 2026.

[137] Abdullah Hosseini and Ahmed Serag. MedToolica: Finetuning-free agentic compositional tool learning for 3D CT reasoning. *Machine Learning and Knowledge Extraction*, 8(6):162, 2026.

[138] Ayhan Can Erdur, Daniel Scholz, Jiazhen Pan, Benedikt Wiestler, Daniel Rueckert, and Jan C. Peeken. Agentic large language models for training-free neuro-radiological image analysis, 2026. URL https://arxiv.org/abs/2604.16729.

[139] Taiping Qu, Hongkai Zhang, Lantian Zhang, Can Zhao, Nan Zhang, Hui Wang, Zhen Zhou, Mingye Zou, Kairui Bo, Pengfei Zhao, et al. BAAI Cardiac Agent: An intelligent multimodal agent for automated reasoning and diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging. *arXiv preprint arXiv:2604.04078*, 2026.

[140] Zhusi Zhong, Yuli Wang, Jing Wu, Wen-Chi Hsu, Vin Somasundaram, Lulu Bi, Shreyas Kulkarni, Zhuoqi Ma, Scott Collins, Grayson Baird, et al. Vision-language model for report generation and outcome prediction in CT pulmonary angiogram. *npj Digital Medicine*, 8(1):432, 2025.

[141] Duaa Alim, Mogtaba Alim, and Liam Chalcroft. GAZE: Grounded agentic zero-shot evaluation with viewer-level tools and literature retrieval on rare brain MRI. *arXiv preprint arXiv:2605.00876*, 2026.

[142] Gulfam Ahmed Sajua, Marjan Akhib, and Yuchou Chang. AgentMRI: A vison language model-powered AI system for self-regulating MRI reconstruction with multiple degradations. *Journal of Imaging Informatics in Medicine*, 39(2):1422–1440, 2025.

[143] Lujia Zhong, Yihao Xia, Jianwei Zhang, Shuo Huang, Jiaxin Yue, Mingyang Xia, and Yonggang Shi. NeuroAgent: LLM agents for multimodal neuroimaging analysis and research. *arXiv preprint arXiv:2605.06584*, 2026.

[144] Keqi Han, Xiang Li, Songlin Zhao, Yao Su, Yixuan Yuan, Lifang He, and Carl Yang. Towards a virtual neuroscientist: Autonomous neuroimaging analysis via multi-agent collaboration. *arXiv preprint arXiv:2605.09366*, 2026.

[145] Andrew Hoopes, Neel Dey, Victor Ion Butoi, John V Guttag, and Adrian V Dalca. VoxelPrompt: A vision agent for end-to-end medical image analysis. *arXiv preprint arXiv:2410.08397*, 2024.

[146] Shengyuan Liu, Liuxin Bao, Qi Yang, Wanting Geng, Boyun Zheng, Chenxin Li, Wenting Chen, Houwen Peng, and Yixuan Yuan. MedSAM-Agent: Empowering interactive medical image segmentation with multi-turn agentic reinforcement learning. *arXiv preprint arXiv:2602.03320*, 2026.

[147] Ziyan Huang, Haoyu Wang, Jin Ye, Yuanfeng Ji, Xiaowei Hu, Lihao Liu, Zhikai Yang, Wei Li, Ming Hu, Yanzhou Su, et al. MedSegAgent: A universal and scalable multi-agent system for instructive medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, pages 1–12, 2026.

[148] Aizierjiang Aiersilan and Mohamad Koubeissi. Neuro-Oracle: A trajectory-aware agentic RAG framework for interpretable epilepsy surgical prognosis. *arXiv preprint arXiv:2604.14216*, 2026.

[149] Rajat Vahistha, Sandra Brosda, Lauren G Aoude, Jessica Ng, Parveen Kundu, Andrew P Barbour, and Viktor Vegh. Agent-MIRA: AI-orchestrated medical imaging agent for PET image retrieval and assistance. *Computerized Medical Imaging and Graphics*, 125:102725, 2026.

[150] Zhihao Chen, Jiahui Wang, Yizhou Chen, Xiaozhong Ji, Xiaobin Hu, Jimin Hong, Wolfram Andreas Bosbach, Axel Rominger, Ali Afshar-Oromieh, Hongming Shan, et al. TheraAgent: Multi-agent framework with self-evolving memory and evidence-calibrated reasoning for PET theranostics. *arXiv preprint arXiv:2603.13676*, 2026.

[151] Mohamed Sobhi Jabal, Jikai Zhang, Dominic LaBella, Jessica L Houk, Dylan Zhang, Jeffrey D Rudie, Kirti Magudia, Maciej A Mazurowski, and Evan Calabrese. Agentic automation of BT-RADS scoring: End-to-end multi-agent system for standardized brain tumor follow-up assessment. *arXiv preprint arXiv:2603.21494*, 2026.

[152] Cheng Yang, Hui Jin, Xinlei Yu, Zhipeng Wang, Yaoqun Liu, Fenglei Fan, Dajiang Lei, Gangyong Jia, Changmiao Wang, and Ruiquan Ge. LungNoduleAgent: A collaborative multi-agent system for precision diagnosis of lung nodules. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(35):29793–29801, 2026.

[153] Yi Lin, Yihao Ding, Yonghui Wu, and Yifan Peng. MARCH: Multi-agent radiology clinical hierarchy for CT report generation. *arXiv preprint arXiv:2604.16175*, 2026.

[154] Zhiping Xiao, Junwei Yang, Gongbo Sun, Han Zhang, Hanwen Xu, Yi Yao, Zachary D. Miller, William E. King, Mohammed M. Kanani, Jalal B. Andre, Sammy Chu, Ming Zhang, Paul E. Kinahan, Nathan M. Cross, and Sheng Wang. A multi-agent system for spine MRI report generation from multi-sequence imaging. *arXiv preprint arXiv:2606.08897*, 2026.

[155] Eleftherios Tzanis, Michail E Klontzas, and Antonios Tzortzakakis. DosimeTron: Automating personalized Monte Carlo radiation dosimetry in PET/CT with agentic AI. *arXiv preprint arXiv:2604.06280*, 2026.

[156] Qingxin Wang, Zhongqiu Wang, Minghua Li, Xinye Ni, Rong Tan, Wenwen Zhang, Maitudi Wubulaishan, Wei Wang, Zhiyong Yuan, Zhen Zhang, et al. A feasibility study of automating radiotherapy planning with large language model agents. *Physics in Medicine & Biology*, 70(7):075007, 2025.

[157] Humza Nusrat, Bing Luo, Ryan Hall, Joshua Kim, Hassan Bagher-Ebadian, Anthony Doemer, Benjamin Movsas, and Kundan Thind. Autonomous radiotherapy treatment planning using DOLA: A privacy-preserving, LLM-based optimization agent. *arXiv preprint arXiv:2503.17553*, 2025.

[158] Dongzhao Wang, Zeyun Hu, Yang Li, Dachuan Xu, Ruijie Yang, and Changjing Zhuge. MARTP: a multi-agent simulation framework for automated radiation therapy planning based on LLMs. *Physics in Medicine & Biology*, 71(12):125037, 2026.

[159] Humza Nusrat, Luke Francisco, Bing Luo, Hassan Bagher-Ebadian, Joshua Kim, Karen Chin-Snyder, Salim Siddiqui, Mira Shah, Eric Mellon, Mohammad Ghassemi, et al. Automated stereotactic radiosurgery planning using a human-in-the-loop reasoning large language model agent. *arXiv preprint arXiv:2512.20586*, 2025.

[160] Xingjian Kang, Linda Vorberg, Andreas Maier, Alexander Katzmann, and Oliver Taubmann. Scan-Do Attitude: Towards autonomous CT protocol management using a large language model agent. In *International Workshop on Agentic AI for Medicine*, pages 46–54. Springer, 2025.

[161] Hongyoon Choi, Sungwoo Bae, and Kwon Joong Na. End-to-end PET/CT interpretation and quantification with an LLM-orchestrated AI agent: A real-world pilot study. *Journal of Nuclear Medicine*, page jnumed.126.272362, 2026.

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

[163] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(3): 036501–036501, 2018.

[164] 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. In *Advances in Neural Information Processing Systems*, volume 35, pages 36722–36732, 2022.

[165] 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.

[166] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al. AbdomenAtlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, 97:103285, 2024.

[167] Tugba Akinci D'Antonoli, Lucas K Berger, Ashraya K Indrakanti, Nathan Vishwanathan, Jakob Weiß, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, et al. TotalSegmentator MRI: robust sequence-independent segmentation of multiple anatomic structures in MRI. *arXiv preprint arXiv:2405.19492*, 2024.

[168] Che Liu, Zhongwei Wan, Yuqi Wang, Hui Shen, Haozhe Wang, Kangyu Zheng, Mi Zhang, and Rossella Arcucci. Argus: Benchmarking and enhancing vision-language models for 3D radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16448–16460, 2025.

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

[170] Jiayu Lei, Xiaoman Zhang, Chaoyi Wu, Lisong Dai, Ya Zhang, Yanyong Zhang, Yanfeng Wang, Weidi Xie, and Yuehua Li. AutoRG-Brain: Grounded report generation for brain MRI. *arXiv preprint arXiv:2407.16684*, 2024.

[171] Yixiong Chen, Wenjie Xiao, Pedro RAS Bassi, Xinze Zhou, Sezgin Er, Ibrahim Ethem Hamamci, Zongwei Zhou, and Alan Yuille. Are vision language models ready for clinical diagnosis? a 3D medical benchmark for tumor-centric visual question answering. *arXiv preprint arXiv:2505.18915*, 2025.

[172] Cosmin I Bercea, Jun Li, Philipp Raffler, Evamaria O Riedel, Lena Schmitzer, Angela Kurz, Felix Bitzer, Paula Roßmüller, Julian Canisius, Mirjam L Beyrle, et al. Nova: A benchmark for anomaly localization and clinical reasoning in brain MRI. *arXiv preprint arXiv:2505.14064*, 2025.

[173] Mohammed Baharoon, Luyang Luo, Michael Moritz, Abhinav Kumar, Sung Eun Kim, Xiaoman Zhang, Miao Zhu, Mahmoud H Alabbad, Maha S Alhazmi, Neel P Mistry, et al. ReXGroundingCT: A 3D chest CT dataset for segmentation of findings from free-text reports. *NEJM AI*, 3(7):AIdbp2501220, 2026.

[174] Tien Nguyen, Dac Nguyen, Trung Thanh Nguyen, Truong Thao Nguyen, Hieu Pham, Johan Barthelemy, Tran Minh Quan, Quoc Viet Hung Nguyen, Thanh Tam Nguyen, Mai Son, et al. Toward a vision-language foundation model for medical data: Multimodal dataset and benchmarks for Vietnamese PET/CT report generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[175] Lama Moukheiber, Caleb M Yeung, Haotian Xue, Alec Helbling, Zelin Zhao, and Yongxin Chen. Beyond a single frame: Multi-frame spatially grounded reasoning across volumetric MRI. *arXiv preprint arXiv:2604.15808*, 2026.

[176] Quoc-Huy Trinh, Xi Ding, Yang Liu, Zhenyue Qin, Xingjian Li, Gorkem Durak, Halil Ertugrul Aktas, Elif Keles, Ulas Bagci, and Min Xu. Beyond medical diagnostics: How medical multimodal large language models think in space. *arXiv preprint arXiv:2603.13800*, 2026.

[177] Mingyang Yu, Yaozong Gao, Yiran Shu, Yanbo Chen, Jingyu Liu, Caiwen Jiang, Kaicong Sun, Zhiming Cui, Weifang Zhang, Yiqiang Zhan, et al. Location-guided automated lesion captioning in whole-body PET/CT images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 348–357. Springer, 2025.

[178] Sadaf Khademi, Mehran Shabanpour, Reza Taleei, Anastasia Oikonomou, and Arash Mohammadi. AutoRad-Lung: A radiomic-guided prompting autoregressive vision-language model for lung nodule malignancy prediction. *arXiv preprint arXiv:2503.20662*, 2025.

[179] Zhaoyang Jiang, Zhizhong Fu, David McAllister, Yunsoo Kim, and Honghan Wu. LoV3D: Grounding cognitive prognosis reasoning in longitudinal 3D brain MRI via regional volume assessments. *arXiv preprint arXiv:2603.12071*, 2026.

[180] Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Curtis Langlotz, Matthew Lungren, Serena Yeung, Nigam Shah, and Jason Fries. INSPECT: A multimodal dataset for patient outcome prediction of pulmonary embolisms. In *Advances in Neural Information Processing Systems*, volume 36, pages 17742–17772, 2023.

[181] Vincent Andrearczyk, Valentin Oreiller, Moamen Abobakr, Azadeh Akhavanallaf, Panagiotis Balermpas, Sarah Boughdad, Leo Capriotti, Joel Castelli, Catherine Cheze Le Rest, Pierre Decazes, et al. Overview of the HECKTOR challenge at MICCAI 2022: Automatic head and neck tumor segmentation and outcome prediction in PET/CT. In *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, pages 1–30. Springer, 2022.

[182] Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medRxiv*, pages 2019–12, 2019.

[183] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

[184] Yuanzhe Li, Hao Chen, Rui Yin, Juyan Ba, Yu Zhang, and Sheng Lu. Gastric-X: A multimodal multi-phase benchmark dataset for advancing vision-language models in gastric cancer analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2490–2501, 2026.

[185] Idan Tankel, Nir Mazor, Rafi Brada, Christina LeBedis, and Guy Ben-Yosef. INFORM-CT: Integrating LLMs and VLMs for incidental findings management in abdominal CT. *arXiv preprint arXiv:2512.14732*, 2025.

[186] Moritz Zaiss, Amr Aly, Jonathan Endres, Tobias Dornstetter, Simon Weinmüller, and Andreas Maier. Agentic MR sequence development: leveraging LLMs with MR skills for automatic physics-informed sequence development. *arXiv preprint arXiv:2604.13282*, 2026.

[187] Xiangde Luo, Zihan Li, Shaoting Zhang, Wenjun Liao, and Guotai Wang. Rethinking abdominal organ segmentation (RAOS) in the clinical scenario: A robustness evaluation benchmark with challenging cases. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 531–541. Springer, 2024.

[188] Gašper Podobnik, Bulat Ibragimov, Elias Tappeiner, Chanwoong Lee, Jin Sung Kim, Zacharia Mesbah, Romain Modzelewski, Yihao Ma, Fan Yang, Mikołaj Rudecki, et al. HaN-Seg: The head and neck organ-at-risk CT and MR segmentation challenge. *Radiotherapy and Oncology*, 198:110410, 2024.

[189] Carlos E Cardenas, Abdallah SR Mohamed, Jinzhong Yang, Mark Gooding, Harini Veeraraghavan, Jayashree Kalpathy-Cramer, Sweet Ping Ng, Yao Ding, Jihong Wang, Stephen Y Lai, et al. Head and neck cancer patient images for determining auto-segmentation accuracy in T2-weighted magnetic resonance imaging through expert manual segmentations. *Medical Physics*, 47(5):2317–2322, 2020.

[190] Mashrafi Monon, Umaima Rahman, Asif Hanif, Numan Saeed, and Mohammad Yaqub. Lost in volume: The CT-SpatialVQA benchmark for evaluating semantic-spatial understanding of 3D medical vision-language models. *arXiv preprint arXiv:2605.08787*, 2026.

[191] Minh Khoi Nguyen, Dai Lam Le, Amir Reza Jafari, Tuan Dung Nguyen, Mai Hong Son, Mai Huy Thong, Quang Huy Nguyen, Thanh Trung Nguyen, Reza Farahbakhsh, Noel Crespi, et al. Med-StepBench: A hierarchical reasoning framework for evaluating hallucinations in medical vision-language models. *arXiv preprint arXiv:2605.10002*, 2026.

[192] Hashmat Shadab Malik, Anees Ur Rehman Hashmi, Numan Saeed, Muzammal Naseer, Salman Khan, and Christoph Lippert. CORTEX: A structured reasoning benchmark for trustworthy 3D chest CT MLLMs. *arXiv preprint arXiv:2606.27264*, 2026.

[193] Yiming Shi, Shaoshuai Yang, Xi Chen, Haolin Li, Hengyu Zhang, Che Jiang, Kaiwen Wang, Xun Zhu, Dong Xie, Fei Wang, et al. ReportQA: QA-based radiology report evaluation. *arXiv preprint arXiv:2606.15037*, 2026.

[194] Bo Liu, Hanxue Gu, Xiangru Li, Zheren Zhu, Jacob Ellison, Kang Wang, Janine M Lupo, Yang Yang, and Hui Lin. Automated report-derived oncology VQA benchmark for evaluating vision-language models on 3D medical imaging. *arXiv preprint arXiv:2606.02809*, 2026.

[195] Bulat Maksudov, Vladislav Kurenkov, Kathleen M Curran, and Alessandra Mileo. ABRA: Agent benchmark for radiology applications. *arXiv preprint arXiv:2605.11224*, 2026.

[196] Tajamul Ashraf, Hyewon Jeong, Fida Mohammad Thoker, and Bernard Ghanem. MedCTA: A benchmark for clinical tool agents. *arXiv preprint arXiv:2606.11702*, 2026.

[197] Qiaoyu Zheng, Chaoyi Wu, Pengcheng Qiu, Lisong Dai, Ya Zhang, Yanfeng Wang, and Weidi Xie. How well can modern LLMs act as agent cores in radiology environments? *arXiv preprint arXiv:2412.09529*, 2024.

[198] Sebastian Wind, Tri-Thien Nguyen, Jeta Sopa, Mahshad Lotfinia, Sebastian Bickelhaup, Michael Uder, Harald Köstler, Gerhard Wellein, Sven Nebelung, Daniel Truhn, et al. Safety and accuracy follow different scaling laws in clinical large language models. *arXiv preprint arXiv:2605.04039*, 2026.

[199] Shansong Wang, Mojtaba Safari, Qiang Li, Chih-Wei Chang, Richard LJ Qiu, Justin Roper, David S Yu, and Xiaofeng Yang. Triad: Vision foundation model for 3D magnetic resonance imaging. *Research Square*, pages rs–3, 2025.

[200] Antoine Saporta, Baptiste Callard, Corentin Dancette, Julien Khlaut, Charles Corbière, Leo Butsanets, Amaury Prat, and Pierre Manceron. Curia-2: Scaling self-supervised learning for radiology foundation models. *arXiv preprint arXiv:2604.01987*, 2026.

[201] Mohammad R Salmanpour, Somayeh Sadat Mehrnia, Sajad Jabarzadeh Ghandilu, Zhino Safahi, Sonya Falahati, Shahram Taeb, Ghazal Mousavi, Mehdi Maghsudi, Ahmad Shariftabrizi, Ilker Hacihaliloglu, et al. Handcrafted vs. deep radiomics vs. fusion vs. deep learning: A comprehensive review of machine learning -based cancer outcome prediction in PET and SPECT imaging. *Journal of Imaging Informatics in Medicine*, pages 1–50, 2026.

[202] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision*, pages 565–571. IEEE, 2016.

[203] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2): 203–211, 2021.

[204] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

[205] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[206] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[207] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

[208] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

[209] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

[210] Mohammad Salmanpour, Shahram Taeb, Ali Fathi Jouzdani, Mohammad Ayazi, Siavash Hosseinpour Saffarian, Mehdi Maghsudi, Ilker Hacihaliloglu, and Arman Rahmim. A clinically anchored radiomics dictionary for explainable TI-RADS–based thyroid nodule classification in ultrasound; dictionary version TU1.0. *European Journal of Radiology*, 203:113014, 2026.

[211] Mohammad R Salmanpour, Sajad Amiri, Sara Gharibi, Ahmad Shariftabrizi, Yixi Xu, William B Weeks, Arman Rahmim, and Ilker Hacihaliloglu. Radiological and biological dictionary of radiomics features: Addressing understandable AI issues in personalized prostate cancer, dictionary version PM1.0. *Journal of Imaging Informatics in Medicine*, 39(3):1929–1950, 2026.

[212] Ali Fathi Jouzdani, Shahram Taeb, Mehdi Maghsudi, Arman Gorji, Arman Rahmim, and Mohammad R Salmanpour. Towards interpretable AI in personalized medicine through a radiological-biological radiomics dictionary linking semantic Lung-RADS and imaging radiomics features. *Journal of Biomedical Informatics*, 179:105047, 2026.

[213] Arman Gorji, Nima Sanati, Amir Hossein Pouria, Somayeh Sadat Mehrnia, Ilker Hacihaliloglu, Arman Rahmim, and Mohammad R Salmanpour. Radiological and biological dictionary of radiomics features: addressing understandable AI issues in personalized breast cancer; dictionary version BM1.0. *Physics in Medicine & Biology*, 71(2):025008, 2026.

[214] Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. MedRAX: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673*, 2025.

[215] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[216] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[217] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139, pages 8748–8763. PMLR, 2021.

[218] Mengzhuo Chen, Yan Shu, Chi Liu, Hongming Piao, Xidong Wang, Derek Li, and Bryan Dai. UniReason-Med: A shared grounded reasoning interface for 2D-to-3D transfer in medical VQA. *arXiv preprint arXiv:2606.11740*, 2026.

[219] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[220] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[221] Yang Xing, Jiong Wu, Savas Ozdemir, Ying Zhang, Yang Yang, Wei Shao, and Kuang Gong. MedVL-SAM2: A unified 3D medical vision-language model for multimodal reasoning and prompt-driven segmentation. *arXiv preprint arXiv:2601.09879*, 2026.

[222] Leilei Zeng, Jie Liu, Wenting Chen, Chenyang Lyu, Wenxi Li, Shaonan Liu, Xiande Zhou, and Linlin Shen. Enhancing 3D medical multi-modal large language models with integrated human body priors for computed tomography. *Pattern Recognition*, 179:113540, 2026.

[223] Chenyu Wang, Weicheng Dai, Han Liu, Wenchao Li, and Kayhan Batmanghelich. Enhancing fine-grained spatial grounding in 3D CT report generation via discriminative guidance. *arXiv preprint arXiv:2604.10437*, 2026.

[224] Pengcheng Shi, Minghui Zhang, Kehan Song, Jiaqi Liu, Yun Gu, and Xinglin Zhang. U-VLM: Hierarchical vision language modeling for report generation. *arXiv preprint arXiv:2603.00479*, 2026.

[225] Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Dong Yang, Pengfei Guo, Marc Edgar, Daguang Xu, Bernhard Kainz, and Bjoern Menze. Better tokens for better 3D: Advancing vision-language modeling in 3D medical imaging. *arXiv preprint arXiv:2510.20639*, 2025.

[226] Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024.

[227] Yan Shu, Chi Liu, Robin Chen, Derek Li, and Bryan Dai. Fleming-VL: Towards universal medical visual reasoning with multimodal LLMs. *arXiv preprint arXiv:2511.00916*, 2025.

[228] Shengyuan Liu, Zanting Ye, Yunrui Lin, Chen Hu, Wanting Geng, Xu Han, Bulat Ibragimov, Yefeng Zheng, and Yixuan Yuan. MedPruner: Training-free hierarchical token pruning for efficient 3D medical image understanding in vision-language models. *arXiv preprint arXiv:2603.11625*, 2026.

[229] Yang Yu, Dunyuan Xu, Yaoqian Li, Xiaomeng Li, Jinpeng Li, and Pheng-Ann Heng. Adapting 2D multi-modal large language model for 3D CT image analysis. *arXiv preprint arXiv:2604.10233*, 2026.

[230] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2GenGPT: Radiology report generation with frozen LLMs. *Meta-Radiology*, 1(3):100033, 2023.

[231] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[232] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. PMC-LLaMA: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, 2024.

[233] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[234] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[235] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The LLaMA 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[236] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[237] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[238] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7543–7557, 2025.

[239] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[240] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[241] Tom Maye-Lasserre, Yitong Li, Bailiang Jian, Morteza Ghahremani, Benedikt Wiestler, and Christian Wachinger. Generating reports or repeating templates? measuring and mitigating template collapse in 3D CT report generation. *arXiv preprint arXiv:2605.30984*, 2026.

[242] Tianwei Lin, Zhongwei Qiu, Jie Cao, Jiang Liu, Wenjie Yan, Bo Zhang, Yu Zhong, Wenqiao Zhang, Yingda Xia, and Ling Zhang. Regulating anatomy-aware rewards via trajectory-integral feedback for volumetric computed tomography analysis. *arXiv preprint arXiv:2605.20277*, 2026.

[243] Anni Tziakouri and Filippo Menolascina. Reinforcement learning for clinical reasoning: Aligning LLMs with ACR imaging appropriateness criteria. *arXiv preprint arXiv:2510.05194*, 2025.

[244] Anglin Liu, Rundong Xue, Xu R Cao, Yifan Shen, Yi Lu, Xiang Li, Qianqian Chen, and Jintai Chen. MedSAM3: Delving into segment anything with medical concepts. *arXiv preprint arXiv:2511.19046*, 2025.

[245] Yankai Jiang, Qiaoru Li, Binlu Xu, Haoran Sun, Chao Ding, Junting Dong, Yuxiang Cai, Xuhong Zhang, and Jianwei Yin. IBISAgent: Reinforcing pixel-level visual reasoning in MLLMs for universal biomedical object referring and segmentation. *arXiv preprint arXiv:2601.03054*, 2026.

[246] Pramit Saha, Joshua Strong, Mohammad Alsharid, Divyanshu Mishra, and J Alison Noble. Picking the right specialist: Attentive neural process-based selection of task-specialized models as tools for agentic healthcare systems. *arXiv preprint arXiv:2602.14901*, 2026.

[247] Zheang Huai, Honglong Yang, and Xiaomeng Li. Which tool response should i trust? tool-expertise-aware chest x-ray agent with multimodal agentic learning. *arXiv preprint arXiv:2602.21517*, 2026.

[248] Satrio Pambudi and Filippo Menolascina. Bridging clinical narratives and ACR appropriateness guidelines: A multi-agent RAG system for medical imaging decisions. *arXiv preprint arXiv:2510.04969*, 2025.

[249] Cheng Wang, Zhibin He, Zhihao Peng, Shengyuan Liu, Yufan Hu, Yang Carl, He Lifang, Lichao Sun, Xiang Li, and Yixuan Yuan. Neuroclaw technical report. *arXiv preprint arXiv:2604.24696*, 2026.

[250] Lianrui Zuo, Yihao Liu, Gaurav Rudravaram, Karthik Ramadass, Aravind R Krishnan, Michael D Phillips, Yelena G Bodien, Mayur B Patel, Paula Trujillo, Yency Forero Martinez, et al. An artifact-based agent framework for adaptive and reproducible medical image processing. *arXiv preprint arXiv:2604.21936*, 2026.

[251] 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.

[252] 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.

[253] 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.

[254] Elizabeth A Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, Danielle Sargent, Robert Ford, Janet Dancey, Stephen Arbuck, Steve Gwyther, Margaret Mooney, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*, 45(2): 228–247, 2009.

[255] Richard L Wahl, Heather Jacene, Yvette Kasamon, and Martin A Lodge. From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumors. *Journal of Nuclear Medicine*, 50(Suppl 1): 122S–150S, 2009.

[256] Patrick Y Wen, David R Macdonald, David A Reardon, Timothy F Cloughesy, A Gregory Sorensen, Evanthia Galanis, John DeGroot, Wolfgang Wick, Mark R Gilbert, Andrew B Lassman, et al. Updated response assessment criteria for high-grade gliomas: Response assessment in neuro-oncology working group. *Journal of Clinical Oncology*, 28(11):1963–1972, 2010.

[257] Abhinav K Jha, Kyle J Myers, Nancy A Obuchowski, Ziping Liu, Md Ashequr Rahman, Babak Saboury, Arman Rahmim, and Barry A Siegel. Objective task-based evaluation of artificial intelligence-based medical imaging methods: framework, strategies, and role of the physician. *PET Clinics*, 16(4):493–511, 2021.

[258] Abhinav K Jha, Tyler J Bradshaw, Irène Buvat, Mathieu Hatt, Prabhat Kc, Chi Liu, Nancy F Obuchowski, Babak Saboury, Piotr J Slomka, John J Sunderland, et al. Nuclear medicine and artificial intelligence: Best practices for evaluation (the RELAINCE guidelines). *Journal of Nuclear Medicine*, 63(9):1288–1299, 2022.

[259] Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, et al. On path to multimodal generalist: General-Level and General-Bench. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 16423–16542, 2025.

[260] Yuqi Hu, Longguang Wang, Xian Liu, Ling-Hao Chen, Yuwei Guo, Yukai Shi, Ce Liu, Anyi Rao, Zeyu Wang, and Hui Xiong. Simulating the real world: A unified survey of multimodal generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2026.

[261] Dyke Ferber, Lars Hilgers, Christiane Höper, Benedict Kinny-Köster, Jan-Niklas Eckardt, Katharina Egger-Heidrich, Marius Bill, Martin MK Schneider, Jan Clusmann, Lejla Kadric, et al. Towards autonomous medical artificial intelligence agents. *Nature*, pages 1–10, 2026.

[262] Khaled Saab, Chunjong Park, Tim Strother, Jan Freyberg, David GT Barrett, Yong Cheng, Wei-Hung Weng, David Stutz, Nenad Tomasev, Anil Palepu, et al. Advancing conversational diagnostic AI with multimodal reasoning. *Nature Medicine*, 32(5):1–11, 2026.

[263] Anil Palepu, Valentin Liévin, Wei-Hung Weng, Khaled Saab, David Stutz, Yong Cheng, Kavita Kulkarni, S Sara Mahdavi, Joëlle Barral, Dale R Webster, et al. Towards conversational AI for disease management. *arXiv preprint arXiv:2503.06074*, 2025.

[264] Nick Woznitza, Lesley Smith, Janette Rawlinson, Iain Au-Yong, Bindu George, Madava G Djearaman, Arjun Nair, Richard W Lee, Neal Navani, Siyabonga Ndwandwe, et al. AI-based chest x-ray prioritization in the lung cancer diagnostic pathway: the LungIMPACT randomized controlled trial. *Nature Medicine*, 32(5):1–8, 2026.

[265] Xiaobin Hu, Yunhang Qian, Jiaquan Yu, Jingjing Liu, Peng Tang, Xiaozhong Ji, Chengming Xu, Jiawei Liu, Xiaoxiao Yan, Xinlei Yu, et al. The landscape of medical agents: A survey. *Authorea Preprints*, 2025.

[266] Chenhui Wang, Boyun Zheng, Liuxin Bao, Zhihao Peng, Peter YM Woo, Hongming Shan, and Yixuan Yuan. Brain-WM: Brain glioblastoma world model. *arXiv preprint arXiv:2603.07562*, 2026.

[267] Tianxingjian Ding, Yuanhao Zou, Chen Chen, Mubarak Shah, and Yu Tian. CLARITY: Medical world model for guiding treatment decisions by modeling context-aware disease trajectories in latent space. *arXiv preprint arXiv:2512.08029*, 2025.

[268] Yuxin Zuo, Zikai Xiao, Li Sheng, Fei Huang, Jianhong Tu, Yuxuan Liu, Tianyi Tang, Xiaomeng Hu, Yang Su, Qingfeng Lan, et al. Qwen-AgentWorld: Language world models for general agents. *arXiv preprint arXiv:2606.24597*, 2026.

[269] Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, et al. Medical world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8319–8329, 2025.

[270] Ke Liu, Mengxuan Li, Yanyi Bao, Tianyun Zhang, Chong Chu, Jiajun Bu, and Haishuai Wang. Medical world models: representing medical states, modelling clinical dynamics and guiding intervention policies. *arXiv preprint arXiv:2606.16721*, 2026.

[271] Tal Ridnik, Dedy Kredo, and Itamar Friedman. Code generation with AlphaCodium: From prompt engineering to flow engineering. *arXiv preprint arXiv:2401.08500*, 2024.

[272] Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2024, 2024.

[273] Zheng Jiang, Heng Guo, Chengyu Fang, Changchen Xiao, Xinyang Hu, Lifeng Sun, and Minfeng Xu. MedVR: Annotation-free medical visual reasoning via agentic reinforcement learning. *arXiv preprint arXiv:2604.08203*, 2026.

[274] Shengyuan Liu, Jia-Xuan Jiang, Boyun Zheng, Cheng Wang, Zipei Wang, Wentao Pan, Hongtao Wu, Houwen Peng, Yu Gu, Lichao Sun, et al. Towards autonomous and auditable medical imaging model development. *arXiv preprint arXiv:2607.10522*, 2026.

[275] Tongrui Zhang, Chenhui Wang, Yongming Li, Zhihao Chen, Xufeng Zhan, and Hongming Shan. Skill-evolving grounded reasoning for free-text promptable 3D medical image segmentation. *arXiv preprint arXiv:2603.08215*, 2026.

[276] Haoran Sun, Wenjie Li, Yujie Zhang, Zekai Lin, Fanrui Zhang, Kaitao Chen, Xingqi He, Yichen Li, Mianxin Liu, Lei Liu, et al. Experience makes skillful: Enabling generalizable medical agent reasoning via self-evolving skill memory. *arXiv preprint arXiv:2606.09365*, 2026.

[277] Timothy Ossowski, Xinchi Liu, Danyal Maqbool, Vaibhav Dhanuka, Sheng Zhang, Hoifung Poon, Majid Afshar, Tyler Bradshaw, and Junjie Hu. CodeClinic: Evaluating automation of coding skills for clinical reasoning agents. *arXiv preprint arXiv:2605.09675*, 2026.

[278] Fanxuan Chen, Haoman Chen, Tao Yu, Ruoyun Wang, Yi Wang, Xian Zhang, Jiachen Li, Kaishuo Liu, Darong Hai, Xueying Bao, et al. AI-driven revolution of medical robotics across surgical innovation, rehabilitation intelligence, and multimodal healthcare delivery. *MedComm*, 7(3):e70597, 2026.

[279] Guankun Wang, Long Bai, and Hongliang Ren. How can reasoning capability empower the AI copilot robot in endoscopic surgery. *npj Digital Medicine*, 9(1):447, 2026.

[280] Ziyang Dong, Xiaomei Wang, Ge Fang, Zhuoliang He, Justin Di-Lang Ho, Chim-Lee Cheung, Wai Lun Tang, Xiaochen Xie, Liyuan Liang, Hing-Chiu Chang, et al. Shape tracking and feedback control of cardiac catheter using MRI-guided robotic platform–validation with pulmonary vein isolation simulator in MRI. *IEEE Transactions on Robotics*, 38(5):2781–2798, 2022.

[281] Alejandro Granados, Raghav Khanna, Nikola Fischer, Nicholas Raison, Margarita Ciabattini, Harry Robertshaw, Maxence Boels, Mohsan Malik, Veronica Granados, Tom Vercauteren, et al. Evolving surgical teams in the age of artificial intelligence and robotics. *Frontiers in Science*, 4:1783803, 2026.

[282] Davy van de Sande, Eline Fung Fen Chung, Jacobien Oosterhoff, Jasper van Bommel, Diederik Gommers, and Michel E van Genderen. To warrant clinical adoption AI models require a multi-faceted implementation evaluation. *npj Digital Medicine*, 7(1):58, 2024.

[283] Beau Norgeot, Giorgio Quer, Brett K Beaulieu-Jones, Ali Torkamani, Raquel Dias, Milena Gianfrancesco, Rima Arnaout, Isaac S Kohane, Suchi Saria, Eric Topol, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature Medicine*, 26(9):1320–1324, 2020.

[284] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, Alastair K Denniston, Hutan Ashrafian, Andrew L Beam, An-Wen Chan, Gary S Collins, Ara Darzi, Jonathan J Deeks, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *The Lancet Digital Health*, 2(10):e537–e548, 2020.

[285] Baptiste Vasey, Myura Nagendran, Bruce Campbell, David A Clifton, Gary S Collins, Spiros Denaxas, Alastair K Denniston, Livia Faes, Bart Geerts, Mudathir Ibrahim, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ*, 377: e070904, 2022.

[286] Ali S Tejani, Michail E Klontzas, Anthony A Gatti, John T Mongan, Linda Moy, Seong Ho Park, Charles E Kahn Jr, and CLAIM 2024 Update Panel. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiology: Artificial Intelligence*, 6(4):e240300, 2024.

[287] Viknesh Sounderajah, Ahmad Guni, Xiaoxuan Liu, Gary S Collins, Alan Karthikesalingam, Sheraz R Markar, Robert M Golub, Alastair K Denniston, Shravya Shetty, David Moher, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature Medicine*, 31(10):3283–3289, 2025.

[288] Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

[289] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[290] Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A Valliani, Jeff Zhang, Gabriel R Rosenbaum, Ashley K Amend-Thomas, David B Kurland, et al. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 31(2):618–626, 2025.

[291] Arjun Mahajan, Ziad Obermeyer, Roxana Daneshjou, Jenna Lester, and Dylan Powell. Cognitive bias in clinical large language models. *npj Digital Medicine*, 8(1):428, 2025.

[292] Ahmad Fayaz-Bakhsh, Janice Tania, Syaheerah Lebai Lutfi, Abhinav K Jha, and Arman Rahmim. What is implementation science: And why it matters for bridging the artificial intelligence innovation-to-application gap in medical imaging. *PET Clinics*, 21(1):1–16, 2026.

[293] Hamid Abdollahi, Ahmad Fayaz-Bakhsh, Ivan Klyuzhin, Madjid Soltani, Babak Saboury, and Arman Rahmim. Computational oncology and the augmented oncologist: How implementation-ready AI and digital twins will transform education, research, and practice in precision oncology—insights from theranostics. *Frontiers in Biomedical Technologies*, 2026. In press.